ConTNet: Why not use convolution and transformer at the same time?

Haotian Yan[1,2*] Zhe Li[2*] Weijian Li[2] Changhu Wang[2] Ming Wu[1] Chuang Zhang[1]
[1]School of AI, Beijing University of Posts and Telecommunications.
{yanhaotian, wuming, zhangchuang}@bupt.edu.cn
[2]ByteDance AI Lab, Beijing.
{lizhe.axiel, liweijian, wangchanghu}@bytedance.com

Abstract

Although convolutional networks (ConvNets) have enjoyed great success in computer vision (CV), it suffers from capturing global information crucial to dense prediction tasks such as object detection and segmentation. In this work, we innovatively propose ConTNet (Convolution-Transformer Network), combining transformer with ConvNet architectures to provide large receptive fields. Unlike the recently-proposed transformer-based models (e.g., ViT, DeiT) that are sensitive to hyper-parameters and extremely dependent on a pile of data augmentations when trained from scratch on a midsize dataset (e.g., ImageNet1k), ConTNet can be optimized like normal ConvNets (e.g., ResNet) and preserve an outstanding robustness. It is also worth pointing that, given identical strong data augmentations, the performance improvement of ConTNet is more remarkable than that of ResNet. We present its superiority and effectiveness on image classification and downstream tasks. For example, our ConTNet achieves 81.8% *top-1 accuracy on ImageNet which is the same as DeiT-B with less than* 40% *computational complexity. ConTNet-M also outperforms ResNet50 as the backbone of both Faster-RCNN (by* 2.6%*) and Mask-RCNN (by* 3.2%*) on COCO2017 dataset. We hope that ConTNet could serve as a useful backbone for CV tasks and bring new ideas for model design. The code will be released at https://github.com/yanhao-tian/ConTNet.*

1. Introduction

Convolutional networks (ConvNets) are widely used in computer vision and become a dominating method in almost all intelligent vision systems [26, 31, 43, 13]. Since most landmark ConvNets mainly use 3x3 convolutions, the receptive field is limited within a local neighbourhood to capture local representation. However, the large receptive filed is of great importance to construct a contextual visual understanding especially in some downstream tasks such as object detection and semantic segmentation. In order to enlarge the receptive field of the ConvNet, stacking multiple convolutional layers (conv layer) seems to be the consensus and induce the flourishing of convolutional backbones useful for various downstream tasks [38, 16, 15, 22].

In Natural Language Processing (NLP), the very major issue is how to model long-range dependencies in long sequences. Self-attention mechanism represented by transformer has become the foremost method and achieved state-of-the-art results on many NLP tasks. The success of transformer [45] strongly motivates researchers to utilize trans-

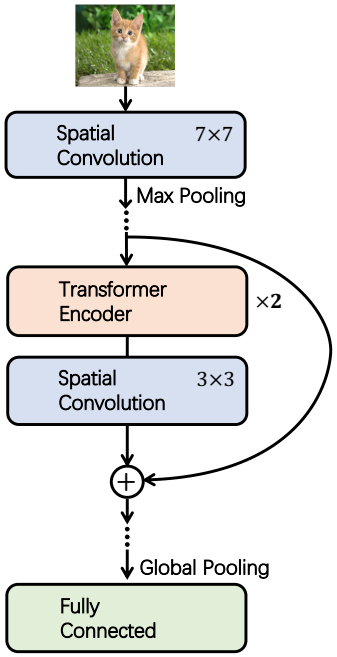

Figure 1. Illustration of the proposed ConTNet framework. ConTNet contains multiple ConT blocks, which are composed of two transformer encoder layers and a convolution layer.

*contributed equally

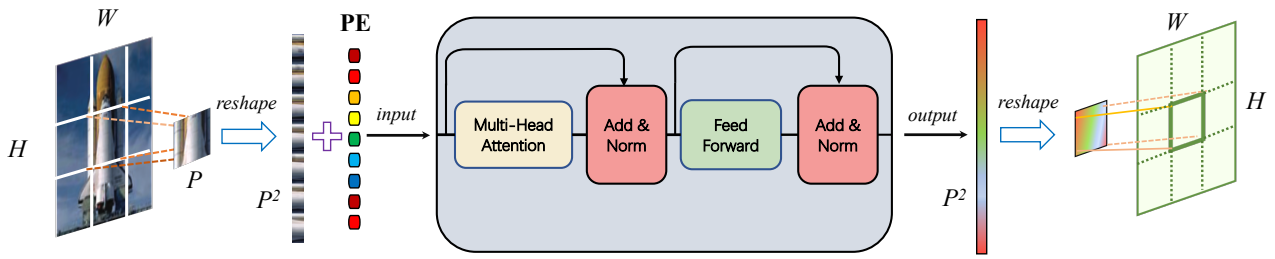

Figure 2. A Patch-wise Standard Transformer Encoder (STE) in ConTNet. **PE** denotes positional encoding. H and W is the height and width of the input and output image separately. P is the size of patch. P^2 is the length of the input and output sequence of STE.

former in CV tasks [4, 11, 44]. However, these vision transformer are highly sensitive to training settings such as learning rate, number of training epochs, optimizer, data augmentation, etc.

We mainly concentrate on the following challenges in CV: (1) ConvNets is deficient in large receptive fields due to the locality of convolution, leading to a performance degradation on downstream tasks. (2) The transformer-based vision model requires special training settings or hundreds of millions of images as the pretraining dataset, which is a practical constraint hampering the widespread adoption.

To overcome these challenges, we propose a novel **Con**volution-**T**ransformer **Net**work (ConTNet) for CV tasks. ConTNet is implemented by stacking multiple ConT blocks as shown in Figure 1. The ConT block treated the standard transformer encoder (STE) as an independent component the same as a conv layer. Specifically, as shown in Figure 2 a feature-map is split into several patches of the same size and each patch flattened to a (super) pixel sequence is next sent to STE. Finally the patch embeddings are reshaped back to feature-maps and they are fed into the next conv layer or STE.

In the proposed ConTNet, the STE plays a leading role in capturing more contextual features, while conv layers efficiently extract local visual information. Besides, we also find that embedding STE into ConvNet architectures can make the network more robust. Or in other words, ConTNet can be trained easily just like the most popular ResNet [15].

We demonstrate that ConTNet is superior to DeiT (transformer-based network) and ResNet (convolution-based network) on ImageNet classification. In accordance with our empirical results, ConTNet can be optimized straightforward like normal ConvNets, and therefore do not requires as many tricks as DeiT [44] or the tremendous amount of pretraining data used by ViT [11]. Another interesting finding is that ConTNet gains more performance improvement from strong data augmentation and other training tricks than that of ResNet [15], which can be attributed to the overfitting risk of transformer architecture. Some key results are listed below: Our ConT-M achieves a 1.6% top-1 accuracy over ResNet50 on ImageNet [9] dataset with almost 25% relative save of computational cost. Likewise, we demonstrate that ConTNet significantly improves dense prediction tasks, especially object detection and segmentation, against the most popular backbone ResNet [15]. Our ConT-M yields around 3% improvement compared to ResNet-50 [15] based on FCOS [42] and Mask-RCNN [14].

In a nutshell, this work's contributions are threefold.

1) To our knowledge, our proposed ConTNet is the first exploration to build a neural network with both of the standard transformer encoder (STE) and spatial convolution.

2) In contrast to the recently trendy visual transformer, ConTNet is much easier to optimize. In addition to robustness, ConTNet performs excellently on image recognition.

3) The empirical results present that a good transfer learning performance is promised. These results suggest that ConTNet provides a new conv-transformer-based pattern to enlarge model's receptive field.

2. Related Work

2.1. ConvNets

Deep Convolutional Neural Networks (ConvNets) enables the computer vision (CV) model to perform at a higher level. This decade, we have witnessed that many novel ideas make ConvNets unprecedented powerful, generalizable, and computationally economical [15, 22, 46, 18, 25, 26, 5, 17, 38, 23]. A succession of researchers improves the performance by deepening the network and adopt the multi-branch architecture that lays the foundation for modern network design [15, 22, 41, 49, 16, 50, 40]. An alternative line of practitioners turns to reformulate convolution or basic block as versatile image features [8, 55, 21, 47, 12]. One of these functional attempts is to incorporate self-attention mechanism into ConvNets, which promotes global features utilization of ConvNets [47, 20, 3, 33, 48, 21]. SENet [21]

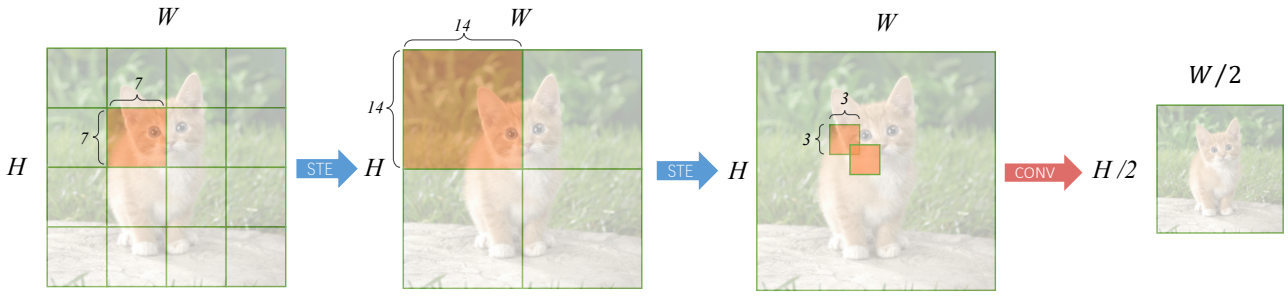

Figure 3. Information flow in a ConT block. For the area covered by orange shadow, the blue arrow indicates that output value of each spatial position is computed from the entire covered area through a patch-wise STE. The red arrow is a conv layer with a kernel size of 3 and a stride of 2, computing only the output value of center position from the entire covered area.

models the interdependencies between channels according to the global context of feature-map. GENet [20] is a generalization of SENet, still first gathering global context as a signal to recalibrate the feature channel. Like CBAM [48] and BAM [33], multiple papers rescale both of different spatial positions and channels by aggregating all contextual information. Non-Local Net [47] implements a pairwise pixel interaction to augment the long-range dependencies across all temporal frames and spatial positions, and Non-local block can introduce the self-attention mechanism. However, the usage of self-attention by Non-Local Net has weaknesses, for instance, the Non-Local Block is hard-weight and insensitive to position of pixels. Another series of works seek to adopt self-attention mechanism along the entire network or even replacing all spatial convolution with self-attention to construct an efficient fully attentional network [35, 19, 2, 52]. These efforts also aim to enhance the long-range dependencies of ConvNets.

Nevertheless, many of these attentional networks are designed elaborately and have not shown strong applicability for downstream tasks. Based on these observations, we innovatively develop ConTNet combining standard transformer encoder (STE) together with convolution layer (conv layer). The network design follows the principle of STE for global features and conv layer for local features. Such an architecture provides a reasonable formula for elevating the network's ability to model long-range dependencies.

2.2. Transformer

Transformer [45] is an encoder-decoder neural network for sequence-to-sequence tasks, which has achieved many state-of-the-art results and further revolutionized NLP with the success of BERT [10]. The recently trendy visual transformer has shown that an end-to-end standard transformer can implement image classification and other vision tasks [4, 30, 24, 54, 56]. ViT [11] cuts the images into some non-overlapping patches and encodes the patches set as a token sequence, whose head is attached to a learnable classification token. The performance of ViT depends on large-scale pretrain datasets like ImageNet-21k dataset or JFT-300M dataset, which trumps the convolutional inductive bias. DeiT [44] explores distillation to extend ViT to a data-efficient vision transformer straightly trained on ImageNet, but the training course of DeiT is complex and unstable. Transformer has also been extended to solving dense prediction problems or low-level tasks. For example, DETR [4] is the first work to using transformer for object detection. DETR uses ConvNets to extract features and uses transformer to model the object detection as an end-to-end dictionary lookup problem. However, DETR is still very sensitive to the hyper-parameters setting and requires a longer training period.

In our practice, embedding STEs into ConvNet, which seems to alternately employ transformer and convolution, can make transformer architecture as robust as convolution (see Figure 2). We combine transformer and convolution as a ConT block, and ConTNet is formed by stacking ConT blocks as shown in Figure 1. ConTNet can be trained from scratch on ImageNet dataset and generalized to dense prediction tasks without unusual settings as expected.

3. ConTNet

In this section, we describe the proposed **Con**volution-**T**ransformer **Net**work (ConTNet) in details.

3.1. Network architecture

We propose the novel convolution-transformer-based network ConTNet. The structure of ConTNet is shown in Figure 1. ConTNet is composed of standard transformer encoders (STEs) and spatial convolutions that are stacked seemingly alternately. More precisely, our very first step towards ConTNet is setting up a relatively shallow ConvNet, which has four stages processing feature-maps with different sizes and channels. This ConvNet is then extended to ConTNet through inserting STE between two neighboring convolutional (conv) layers. Such an extension is made to

Stage	Input size	ConT-Ti	ConT-S	ConT-M	ConT-B
Stage 0	224×224	7×7, 64, stride=2, padding=3			
Stage 1	56×56	$\begin{bmatrix} D=48, \\ D_{ffn}=192, \\ H=1 \end{bmatrix} \times 1$	$\begin{bmatrix} D=64, \\ D_{ffn}=256, \\ H=1 \end{bmatrix} \times 1$	$\begin{bmatrix} D=64, \\ D_{ffn}=256, \\ H=1 \end{bmatrix} \times 2$	$\begin{bmatrix} D=64, \\ D_{ffn}=256, \\ H=1 \end{bmatrix} \times 3$
Stage 2	28×28	$\begin{bmatrix} D=96, \\ D_{ffn}=384, \\ H=2 \end{bmatrix} \times 1$	$\begin{bmatrix} D=128, \\ D_{ffn}=512, \\ H=2 \end{bmatrix} \times 1$	$\begin{bmatrix} D=128, \\ D_{ffn}=512, \\ H=2 \end{bmatrix} \times 2$	$\begin{bmatrix} D=128, \\ D_{ffn}=512, \\ H=2 \end{bmatrix} \times 4$
Stage 3	14×14	$\begin{bmatrix} D=192, \\ D_{ffn}=768, \\ H=4 \end{bmatrix} \times 1$	$\begin{bmatrix} D=256\ , \\ D_{ffn}=1024, \\ H=4 \end{bmatrix} \times 1$	$\begin{bmatrix} D=256, \\ D_{ffn}=1024\ , \\ H=4 \end{bmatrix} \times 2$	$\begin{bmatrix} D=256, \\ D_{ffn}=1024, \\ H=4 \end{bmatrix} \times 6$
Stage 4	7×7	$\begin{bmatrix} D=384\ , \\ D_{ffn}=768, \\ H=8 \end{bmatrix} \times 1$	$\begin{bmatrix} D=512\ , \\ D_{ffn}=1024, \\ H=8 \end{bmatrix} \times 1$	$\begin{bmatrix} D=512, \\ D_{ffn}=1024, \\ H=8 \end{bmatrix} \times 2$	$\begin{bmatrix} D=512\ , \\ D_{ffn}=1024, \\ H=8 \end{bmatrix} \times 3$
	7×7	global average pooling, 1000-d fc, softmax			

Table 1. Detailed settings of ConTNet series. Inside the brackets, we list the hyper-parameter of each ConTBlock. D is the embedding dimension of MHSA, D_{ffn} is the dimension of FFN, and H is the head number of MHSA. Outside the brackets, the number of stacked blocks on the stage is presented. In each stage, the last conv layer performs downsampling and increases dimension.

capture global features supplementary to local representations learned by conv layers.

To systematically embedding STEs into ConvNet, we design a block make STE fully integrated with conv layers by grouping them in pairs. A ConT block cascades a pair of STEs and a conv layer as shown in Figure 1. Each ConT block comprises two STEs and one conv layer with a kernel size of 3×3. In our implementation, the spatial size of split patches is set to 7 and 14 sequentially. Inspired by ResNet, we construct a shortcut connection to implement the residual learning $y = f(x) + x$, which is widely utilized to improve the performance of ConvNets. ConTNet is still a 4-stage-style hierarchical network because we desire it to be suitable for downstream tasks, especially object detection and segmentation. In each stage, the last conv layer of the stage has a stride of 2 to conduct downsampling and increasing channels. When trained on the ImageNet dataset, the size of each stage's feature-map is $[56, 28, 14, 7]$, and the channel of each stage is determined by scaling the popular setting $[64, 128, 256, 512]$ used by most ConvNets [15]. The head number of multi-head self attention of STE in each stage is set to $[1, 2, 4, 8]$ to keep the channel of single-head attention 64. Note that all conv layers in ConTNet have a kernel of 3×3 except for the top and the bottom conv layer. A conv layer with a kernel of 7×7 and a MaxPooling are placed at the top of the network, which follows the practice of ResNet. And the last conv layer of the last stage has a kernel of 1×1 to save parameters.

To produce architectural variants of ConTNet, we scale the depth and width of ConTNet to reach different computational budgets. Table 1 presents four architectures: ConT-Ti (Tiny), ConT-S (Small), ConT-M (Medium), ConT-B (Big). From ConT-Ti to ConT-B, the parameters and FLOPs are gradually increased, and the depth or width grows progressively.

3.2. Patch-wise Standard Transformer Encoder

In this subsection, we revisit the Standard Transformer Encoder (STE) and present a procedure for the execution of its capturing long-range dependency in ConTNet.

In the raw application of standard Transformer [45], a sequence of words is received as input and finally translated into a new sentence. When applied to Computer Vision (CV) tasks, Transformer encoder has to take a 2D image

as input. On the principle of retaining authentic STE, the input 2D image $\mathbf{x}_{2\mathrm{d}} \in \mathbb{R}^{H\times W\times C}$ is flattened into a sequence of pixels denoted by $\mathbf{x}_{1\mathrm{d}} \in \mathbb{R}^{(H\times W)\times C}$, where H and W are height and width of the input image separately, and C is the image channel. Towards a simple but powerful utilization of STE, we develop a method that roughly similar to convolutional filter (see Figure 2). Given a 2D-image $\mathbf{x}_{2\mathrm{d}} \in \mathbb{R}^{H\times W\times C_{in}}$, a convolutional filter aggregates spatial information across a local neighbourhood. For example, a kernel of shape $3 \times 3 \times C_{in}$ slides over every position of image to do a inner product of local window and the kernel weights. Assuming that the input $\mathbf{x}_{2\mathrm{d}}$ and the output $\mathbf{y}_{2\mathrm{d}} \in \mathbb{R}^{H\times W\times C_{out}}$ has the same spatial size, for a spatial position $x_{ij;i\in[0,H),j\in[0,W)}$, the corresponding output value $y_{ij;i\in[0,H),j\in[0,W)}$ is calculated as follow:

$$y_{ij} = \mathrm{Conv}\left(x_{ij}\right), \tag{1}$$

$$\mathrm{Conv}\left(x_{ij}\right) = \sum_{a,b=-k}^{k} W_{a,b}\, x_{i+a,j+b}, \tag{2}$$

where $W \in \mathbb{R}^{(2k+1)\times(2k+1)\times C_{in}\times C_{out}}$ is the weight of convlutional filter, and $a, b \in [-k, k]$. With above analysis, a pixel-to-pixel mapping is supported explicitly in convolutional filter. By contrast, the STE learns a mapping from sequence to sequence, as shown in Figure 2. Therefore, a potential way to exploit the STE in a conv layer fashion is considering a sequence as a pixel. Prior to performing on an image, STE first splits it into several patches of the same size $P \times P$, and treats each patch as a sequence of pixels. The split images can be denoted by a new tensor $\mathbf{x}^{\mathrm{p}}_{2\mathrm{d}} \in \mathbb{R}^{H_p\times W_p\times P^2\times C}$, where H_p is set to H/p and W_p is set to W/p. Instead of that each spatial position of $\mathbf{x}_{2\mathrm{d}}$ is a pixel, each spatial position of $\mathbf{x}^{\mathrm{p}}_{2\mathrm{d}}$ is a sequence $x^{p}_{mn;m\in[0,H_p),n\in[0,W_p)} \in \mathbb{R}^{P^2\times C}$. Assuming that the input $\mathbf{x}^{\mathrm{p}}_{2\mathrm{d}}$ has the same spatial size and channel as the output $\mathbf{y}^{\mathrm{p}}_{2\mathrm{d}}$, for a spatial position (m, n), the output value $y^{p}_{mn;m\in[0,H_p),n\in[0,W_p)} \in \mathbb{R}^{P^2\times C}$ is calculated as follow:

$$y^{p}_{mn} = \mathrm{STE}\left(x^{p}_{mn}\right). \tag{3}$$

The operation of STE is enumerated exactly according to [45], which can be formulated:

$$\mathrm{STE}\left(x^{p}_{mn}\right) = \mathrm{FFN}\left(\mathrm{MHSA}\left(x^{p}_{mn} + \mathbf{PE}\right)\right), \tag{4}$$

where $\mathrm{MHSA}\left(\right)$ is a Multi-Head Self Attention (MHSA) mechanism and $\mathrm{FFN}\left(\right)$ is a 2-layer Feed Forward Network (FFN), and $\mathbf{PE} \in \mathbb{R}^{P^2\times C}$ is positional encoding.

Finally, we describe a principled execution of MHSA and FFN. Take a sequence $\mathbf{x}_{1\mathrm{d}} \in \mathbb{R}^{N\times C}$ as input, a single-head attention value is computed using:

$$A = \mathrm{softmax}\left(\frac{(\mathbf{W}_{\mathrm{q}}\mathbf{x}_{1\mathrm{d}})(\mathbf{W}_{\mathrm{k}}\mathbf{x}_{1\mathrm{d}})^{\mathrm{T}}}{\sqrt{D_{\mathrm{h}}}}\right)(\mathbf{W}_{\mathrm{v}}\mathbf{x}_{1\mathrm{d}}), \tag{5}$$

where $\mathbf{W}_{\mathrm{q}}$, $\mathbf{W}_{\mathrm{k}}$ and $\mathbf{W}_{\mathrm{v}} \in \mathbb{R}^{C\times D_h}$ are the learned linear transformations, and D_h is typically set to D/h. D is the embedding dimension and h is the number of head. Multi-head attention value is obtained by projecting concatenated single attention values:

$$A_{mh} = [A_1; A_2; ...; A_h]\,\mathbf{W}_{\mathrm{mhsa}}, \tag{6}$$

$$\mathrm{MHSA}\left(\mathbf{x}_{1\mathrm{d}}\right) = \mathrm{LN}\left(A_{mh} + \mathbf{x}_{1\mathrm{d}}\right), \tag{7}$$

where $\mathbf{W}_{\mathrm{mhsa}} \in \mathbb{R}^{D\times C}$ is the learned weights that aggregates multiple attention values, and $\mathrm{LN}\left()\right.$ is Layernorm [1], applied after a residual connection. The output of MHSA is fed into FFN:

$$\mathrm{FFN}\left(\mathbf{x}_{1\mathrm{d}}\right) = \mathrm{LN}\left(\mathbf{W}_2\mathbf{W}_1\mathbf{x}_{1\mathrm{d}} + \mathbf{x}_{1\mathrm{d}}\right), \tag{8}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C\times D_{ffn}}$ and $\mathbf{W}_2 \in \mathbb{R}^{D_{ffn}\times C}$ are both learned linear transformations.

3.3. Analysis and More Details

We have revealed that (1) the STE is adopted on each sequence $x^{p}_{mn;m\in[0,H_p),n\in[0,W_p)} \in \mathbb{R}^{P^2\times C}$ which is flattened from the uniformly split patch by sliding the window on the separate input image $\mathbf{x}^{\mathrm{p}}_{2\mathrm{d}} \in \mathbb{R}^{H_p\times W_p\times P^2\times C}$ and (2) ConTNet captures both of global and local features by filtering feature-maps with conv layers and STEs alternately. More details about these two implementations are reported in this subsection.

Using STE like a kernel: In ConTNet, an STE slides over the image and translates each split patch into a new patch, which performs like a filter with kernel size and stride both equal to the patch size. We find such an kernel-like operation has two favourable properties. Concretely speaking, the patch-wise STE is weight-shared, which has translation equivariance and computational efficiency. Instead of a pixel-wise translation equivariance, a relatively rough patch-wise translation equivariance is obtained through our operation. For each patch, some work endeavors to model a pixel-wise translation equivariance when performing a self-attention mechanism on a 2D-shape image. To alleviate this issue, we choose a simple but effective method: A parameter-shared learned PE is added to each split patch, recording each pixel's coordinate in a patch and avoid permutation equivariance problem [45]. In contrast with a conv layer with a kernel of 3×3 that has $9C^2$ parameters and a computational complexity of $9C^2HW$ (assuming that the input and the output are of the same channel) , an STE has $2D_{mhsa}D_{ffn} + 4D^2_{mhsa} + P^2D_{mhsa}$ parameters and a computational complexity of $2D_{mhsa}D_{ffn}HW + 4D^2_{mhsa}HW + (HW)/P^2$. When the dimension of MHSA

D_{mhsa} and the input's channel C_{in} are equal and the dimension of FFN D_{ffn} is equal to four times D_{mhsa} (practice of [45]), the increase of parameters is $3C^2 + P^2C$ and the increase of computational compexity is $3C^2HW + (HW)/P^2$.

Alternately capturing features with STE and conv layer: ConvNets are built upon multiple conv layers with an advantage of locality as well as a lack of global features. To take full advantage of Transformer to enlarge receptive fields, we build ConT block, the basic block of ConTNet, containing two STEs and one conv layer. In each ConT block, STEs first capture more global features, and then a conv layer with a small kernel captures more local features. Hence, ConTNet interweaves the features alternately captured by STEs and conv layers via stacking ConT blocks. Owing to the patch-wise operation on the STE, one pixel has interactions with all pixels in its patch, and we hypothesize the size of the split patch can be adjusted flexibly to model kernels of different receptive fields. We introduce a dynamic setting of patch size rather than a fixed version along the entire network. The default patch size of the first and the second STE is set to 7 and 14 separately, and the kernel size of conv layer is set to 3 (see Figure 3). An identity shortcut is connected between the input and output of the ConT block by doing an element-wise addition. To conduct downsampling and dimension changing, the conv layer of the last ConT block in each stage has a stride of 2 and increases dimension. In this special case, a projection shortcut replaces the identity shortcut, using a 1×1 conv layer with a stride of 2 to match the increased dimension and smaller size for element-wise addition.

4. Experiments

In this section, we conduct extensive experiments on image classification and downstream tasks to evaluate the effectiveness of our ConTNet.

4.1. Image Classification

We compare ConTNet to advanced ConvNets (ResNet [15]) and transformer-based backbone (DeiT [44]) separately to assess the image classification performance on ImageNet dataset [9]. ImageNet dataset contains a train set of 1.28 million images and a validation set of 50000 images. We train all models on the train set and test performance on the validation set. ResNet is the chosen competitor against ConTNet since it is the most popular ConvNet in CV applications. On the other hand, we consider DeiT as representative of the visual transformer, which can be directly trained from scratch on ImageNet dataset.

Towards a fair competition, we trained ConTNet and ResNet with identical training settings. We use SGD [39] to train both of ConTNet and ResNet with a batch-size of 512 for 200 epochs on 8 Nvidia V100 GPUs. Cosine decay schedule is used to adjust learning rate. The initial learning rate is set to 0.2 for ResNet and conv layers in ConTNet, and the initial learning rate of STEs in ConTNet is uniquely set to 0.005. We use a weight decay of 0.00005 and a momentum of 0.9. We follow a simple data-augmentation in ResNet and perform regularization by using label smooth with ϵ of 0.1. Table 2 shows the results of comparison. Specifically, ConT-S ourperforms ResNet-18 (by 3.4%), and ConT-M performs sligtly better than ResNet-50 (by 0.5%), while saving nearly 20% parameters as well as 25% computational costs. However, ConT-B achieves a slightly

Network	FLOPs(G)	#Param(M)	Top-1(%)
Res-18	1.8	11.7	71.5
ConT-S	1.5	10.1	**74.9**
Res-50	4.0	25.6	77.1
ConT-M	3.1	19.2	**77.6**
Res-101	7.6	44.5	**78.2**
ConT-B	6.4	39.6	77.9

Table 2. Top1 accuracy $(\%)$ on ImageNet. ResNet and ConTNet are trained with the same setting.

Network	FLOPs(G)	#Param(M)	Top-1(%)
Res-18*	1.8	11.7	73.2
ConT-S*	1.5	10.1	**76.5**
Res-50*	4.0	25.6	78.6
ConT-M*	3.1	19.2	**80.2**
Res-101*	7.6	44.5	80.0
ConT-B*	6.4	39.6	**81.8**

Table 3. Top1 accuracy $(\%)$ on ImageNet. $*$ indicates that the model is trained with strong data augmentation containing mixup and auto-augmentation. AdamW is used to optimize ConTNet here instead of SGD. The fairness of comparsion is enabled by the fact that SGD is fairly good for ResNet.

Network	FLOPs(G)	#Param(M)	Top-1(%)
DeiT-Ti	1.3	5.7	72.2
ConT-Ti*	0.8	5.8	**74.9**
DeiT-S	4.6	22.1	79.8
ConT-M*	3.1	19.2	**80.2**
DeiT-B	17.6	86.6	**81.8**
ConT-B*	6.4	39.6	**81.8**

Table 4. Top1 accuracy $(\%)$ on ImageNet. Results of DeiT is from the original paper. $*$ indicates that model is trained with mixup and auto-augmentation with AdamW optimizer.

Method	Backbone	Param(M)	FLOPs(G)	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
RetinaNet	Res-50	32.0	235.6	36.5	55.4	39.1	20.4	40.3	48.1
	ConT-m	27.0	217.2	37.9(**+1.4**)	58.1(**+2.7**)	40.2(**+1.1**)	23.0(**+2.6**)	40.6(**+0.3**)	50.4(**+2.3**)
FCOS	Res-50	32.2	242.9	36.6	56.0	38.8	21.0	40.6	47.0
	ConT-m	27.2	228.4	39.3(**+2.7**)	59.3(**+3.3**)	41.8(**+3.0**)	23.1(**+2.1**)	43.1(**+2.5**)	51.9(**+4.9**)
Faster RCNN	Res-50	41.5	241.0	37.4	58.1	40.4	21.2	41.0	48.1
	ConT-m	36.6	225.6	40.0(**+2.6**)	62.4(**+4.3**)	43.0(**+2.6**)	25.4(**+4.2**)	43.0(**+1.9**)	52.0(**+3.9**)

Table 5. Object detection mAP (%) on the COCO validation set. We evaluate the performance of detection model by replacing backbone ResNet50 in ReTinaNet, FCOS and Faster-RCNN with ConTNet.

lower accuracy than ResNet-101 (by -0.3%), as the computational complexity is lower. We hypothesize that STE in ConTNet has a high risk of overfitting. To eliminate the risk, we train ConTNet and ResNet both with strong data augmentations including auto-augmentation [7] and mixup [51]. Additionally, we use AdamW [32] optimizer in stead of SGD to train ConTNet with an initial learning rate of 0.0005 and a weight decay of 0.05. Table 3 shows the results with strong augmentations. In the case of using the same strong data augmentation methods, ConTNet earns more benefits than ResNet, and all ConTNets outperform ResNets on ImageNet dataset across different computational budgets. For example, ConT-B achieves 81.8%, which is 1.8% better than Res-101.

The results of DeiT presented in Table 4 is sourced from the original paper, which is trained with a pipeline of strong data augmentations and many tricks with AdamW optimizer [44]. With the same optimizer, fewer data augmentations and other tricks, ConTNet presents a better performance against DeiT. For example, ConT-B achieves 81.8%, the same as DeiT-B, while our FLOPs is 60% fewer (6.4 G $vs.$ 17.6 G).

One can also find that the performance of DeiT is extremely sensitive to the data augmentations and training tricks [44]. For instance, without either stochastic depth or random erasing, the convergence of DeiT is seemingly disabled. If traind with a SGD optimizer instead of AdamW, DeiT achieves a much lower accuracy by 7.3. By contrast, our experiments have shown that ConTNet can be optimized stably even in the same setting as ResNet and only a few data augmentation tricks incur a better accuracy exceeding DeiT, which verifies that ConTNet has an advantage of robustness over DeiT.

4.2. Downstream tasks

Downstream tasks are more sensitive to the size of the receptive field, and models capable of capturing long-range dependencies would achieve better results. To demonstrate ConTNet has a stronger ability to capture global information, we make use of ConT-M as the backbone to implement some downstream tasks, including object detection, instance segmentation and semantic segmentation. Our empirical results show that ConT-M obtains a better performance with lower computational complexity in contrast to using resnet50 as backbone.

Object Detection: We examine the object detection task on the COCO2017 [29] dataset containing a train set of 118k images and a validation set of 5k images. We adopt Faster-RCNN [36] (two-stage), FCOS [42] (one-stage) and RetinaNet [28] (one-stage) as the detection model with ResNet-50 backbone. Both ConT-M and ResNet-50 are pretrained on ImageNet dataset with the same training setting. Then all models are trained through mmdetection following the same training setting. We modify the input size slightly from $(1333, 800)$ to $(1344, 784)$ and employ marginal padding between stage 2 and 3 to match the patch size of STE. Because the ConTNet conducts downsampling with the last conv layer of each stage, we tweak the stride from 2 to 1 and add Average-Pooling layer for downsampling in order to take full use of FPN [27].

Table 5 records the results of detection experiments. The ConTNet-based model outperforms baseline in all object detection methods. Take a close look at these results, AP_l is improved most largely, and AP_s has a higher performance improvement than A_m. This observation proves that ConTNet has a relatively large receptive field to capture more global features without compromising locality. We conclude that using STE and conv layer alternately can learn more contextual representation than ConvNet while maintaining local features. Most notably, such improvements are attained without any bells and whistles, thus providing an evidence that ConTNet is applicable and stable for the downstream task.

Instance Segmentation: We use COCO2017 instance dataset to investigate the performance of ConTNet on instance segmentation. We employ Mask-RCNN [14] with ResNet50 as baseline, following the training settings in object detection experiments. Table 6 shows that ConTNet improves the $bbox_{map}$ by 2.3% and the $segm_{map}$ by

Backbone	AP^{bb}	AP^{bb}_s	AP^{bb}_m	AP^{bb}_l	AP^{mk}	AP^{mk}_s	AP^{mk}_m	AP^{mk}_l
Res-50	38.2	21.9	40.9	49.5	34.7	18.3	37.4	47.2
ConT-M	40.5(**+2.3**)	25.1(**+3.2**)	44.4(**+3.5**)	52.7(**+3.2**)	38.1(**+3.4**)	20.9(**+2.6**)	41.0(**+3.6**)	50.3(**+3.1**)

Table 6. Instance segmentation mAP (%) on the COCO validation set. We replace the ResNet50 of Mask-RCNN with ConT-M.

Model	mIOU(%)
PSP-Res50*	77.30
PSP-Res50	77.12
PSP-ConTM	78.28

Table 7. mIOU on cityscapes validation set. $\star$ indicates the results of original implementation.

3.4%. The experiments on instance segmentation also support the conclusion proposed in the object detection subsection. When ConTNet serves as the backbone, ConTNet improves the model's performance and captures global features better.

Semantic Segmentation: We conduct semantic segmentation experiments on the Cityscapes dataset [6]. We use PSPNet [53] for this implementation. We replace the ResNet50 in PSPNet with ConT-M and follow training and validation protocols in PSPNet except for input size and optimizer. The original input size is$(713, 713)$ and we use a input size of $(672, 672)$ instead to match split operation of STE. We use AdamW optimizer to train our model on training set and evaluate the performance on validation set. Table 7 shows that our model achieves a higher mIOU than baseline (1.16%).

4.3. Ablation study

Positional Encoding: Positional encoding (PE) is essential to Transformer in NLP tasks. The order of words defines the grammar for a sentence, and the order of pixels determines the semantics of an image. In ConTNet, a standard transformer encoder (STE) infuses a shared learnable 2D position encoding into each split patch. Table 8 shows the results of different position encoding schemes. Expectedly, the infusion of PE increases performance of a position-unware ConTNet significantly. Compared to adopting 2D position encoding [34], the performance degrades slightly using 1D position encoding. We next replace self-attention in STE with relative attention from [37] to implement a relative position encoding, which yields an additional improvement.

As mentioned above, ConTNet adopts a patch-wise positional infusion by adding a shared PE to each split patch. We are curious aboout the effect of adding PE to entire image before splitting it into patch. Table 8 shows that patch-wise position encoding outperforms image-wise position encoding (by 0.9%). The result suggests that PE should be used in combination with patch-wise STE more than in isolation.

Position Encoding	Placement	Top-1(%)
None	None	78.9
1D learnable	patch-wise	80.1
2D learnable	patch-wise	80.2
2D learnable	image-wise	79.3
Relative	patch-wise	**80.4**

Table 8. Results of different position encoding schemes on ImageNet classification. In the case of 2D learnable postion encoding, results of different placements are listed.

Patch size	Top-1(%)
[7, 7], [7, 7], ..., [7, 7], [7, 7]	79.7
[14,14], [14,14], ..., [14,14], [14,14]	78.8
[7,7], [14,14], ..., [7,7], [14,14]	80.0
[7,14], [7,14], ..., [7,14], [7,14]	**80.2**

Table 9. Results of different combinations of patch size on ImageNet classfication. Using multiple patch sizes is a better setting.

Impact of different patch sizes: Patch size is an important hyper-parameter in ConTNet. In order to explore the effect of different patch sizes and find a suitable patch size, we test four different arrangements of patch size by tuning ConT-M: 1) All patch size is set to 7, 2) all patch size is to 14, 3) the patch size of the first block in each stage is set to 7 and that of the second block is set to 14, 4) the patch size of first STE in each block is set to 7 and that of second STE is set to 14. Table 9 shows that except setting 2), all patches are 14×14, the results of other arrangements of patch sizes are relatively close. So we choose the patch size of 7 and 14 alternately as the default setting.

Range of learning rate: We would also like to know the impacts of different learning rates on our model. A wide range of learning rate choices means that it can be applied to various downstream tasks and transferred stably. We train ConT-M with SGD optimizer and adjust the learn-

Network	Conv LR	STE LR	Top-1(%)
ConT-M	0.2	0.01	77.4
ConT-M	0.2	0.005	77.6
ConT-M	0.1	0.005	77.6
ConT-M	0.1	0.001	77.1

Table 10. Results of different pairs of learning rates on ImageNet validation. The network is trained with SGD optimizer.

Group number	FLOPs(G)	#Param(M)	Top-1(%)
1	3.1	19.2	80.2
4	2.6	15.0	78.8
8	2.5	14.3	79.3
16	2.4	14.0	79.1
depthwise	2.5	15.3	79.6

Table 11. Results of replacing convolution in ConT Block with group convolution with different number of groups. "depthwise" means that the convolution is depth-wise separable.

ing rate of conv layers and STE separately. In our practice, the initial learning rate of STE is denoted by $lr_{ste} \in \{0.01, 0.005, 0.001\}$ and the initial learning rate of conv layer is denoted by $lr_{conv} \in \{0.2, 0.1\}$. Table 10 shows the results of ConT-M trained with each group of learning rates and we find that the performance of ConT-M is not susceptible to learning rate. Therefore, the performance of applying ConTNet as the backbone of computer vision models can not be limited by the specific learning rate setting adopted in other tasks, which suggests that ConTNet is of great robustness.

Group convolution: ConTNet is equipped with multiple conv layers that aggregate local features supplementary to global information captured by STE. Group convolution is a successful variants of convolution to reduces computational cost. In authentic convolution, the weight of kernel is denoted by $\mathbf{W}_{\mathrm{k}} \in \mathbb{R}^{k\times k\times C_{in}\times C_{out}}$, which is grouped into a set of kernels $\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_{\mathrm{g}} \in \mathbb{R}^{k\times k\times C_{in}/g\times C_{out}/g}$ in a group convolution. The grouped convolutional kernel disentangles the channel number and therefore achieve a lower computational and memory costs than authentic convolution. Table 11 shows the impact of different group settings on performance of ConT-M, which reveals that grouped convolution induces a non-negligible performance degradation for ConTNet. An alternative solution to increase efficiency of our model is to exploit depth-wise separable convolution. As can be seen from Table 11, depth-wise separable convolution make the best trade-off between the computational cost and inference performance.

4.4. Visualization

Visualization of segmentation: We list some visualized results of semantic segmentation to compare performance qualitatively as shown in Figure 4. In the first column, the true class of the area inside the yellow box is labeled as bus both in ground truth and the fourth row. However, in the third row, these pixels are labeled as truck. In the second column, inside the left yellow box is a truck as shown in the second and fourth tow, but in the third row, some pixels of the truck are labeled as car. As for the right yellow box in the second column, the results of the fourth row labeled more pixels than that of the third row. In the third column, a large area of side walk is left unlabeled due to the occlusion of shadow as shown in the yellow box. In contrast, as can be seen from the fourth row, almost all of the sidewalk pixels that were missed in the third row are labeled.

The segmentation results indicated that ConTNet aggregates more global information than ConvNets and the pixels of a large object enjoy a better inner consistency. Such a promotion of segmentation network is stemmed from the ability of modeling long-range dependencies by exploiting STE in ConTNet.

5. Concolusion

In this work, we propose an innovative convolution-transformer network (ConTNet). ConTNet combines transformer and convolution by stacking standard transformer encoder (STE) and spatial convolution alternately. We introduce a weight-shared patch-wise STE to model a large kernel. A series of experiments demonstrate that ConTNet outperforms ResNet and achieves comparable accuracy of trendy visual transformer with much lower computational complexity. We verify that the effectiveness of implementing downstream tasks with ConTNet backbone is primarily derived from aggregating more contextual information. Moreover, we demonstrate that ConTNet is as robust as ConNets for example ResNet. In other words, ConTNet's performance does not rely on strong data augmentations and fancy training tricks, and is not sensitive to hyper-parameters as well. Finally, in ConTNet we use the vanilla convolution and transformer, if replaced with the improved versions of convolution and transformer recently, a further performance is promised. We hope that our work could bring some new ideas for model design.

References

[1] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[2] 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*, pages 3286–3295, 2019. 3

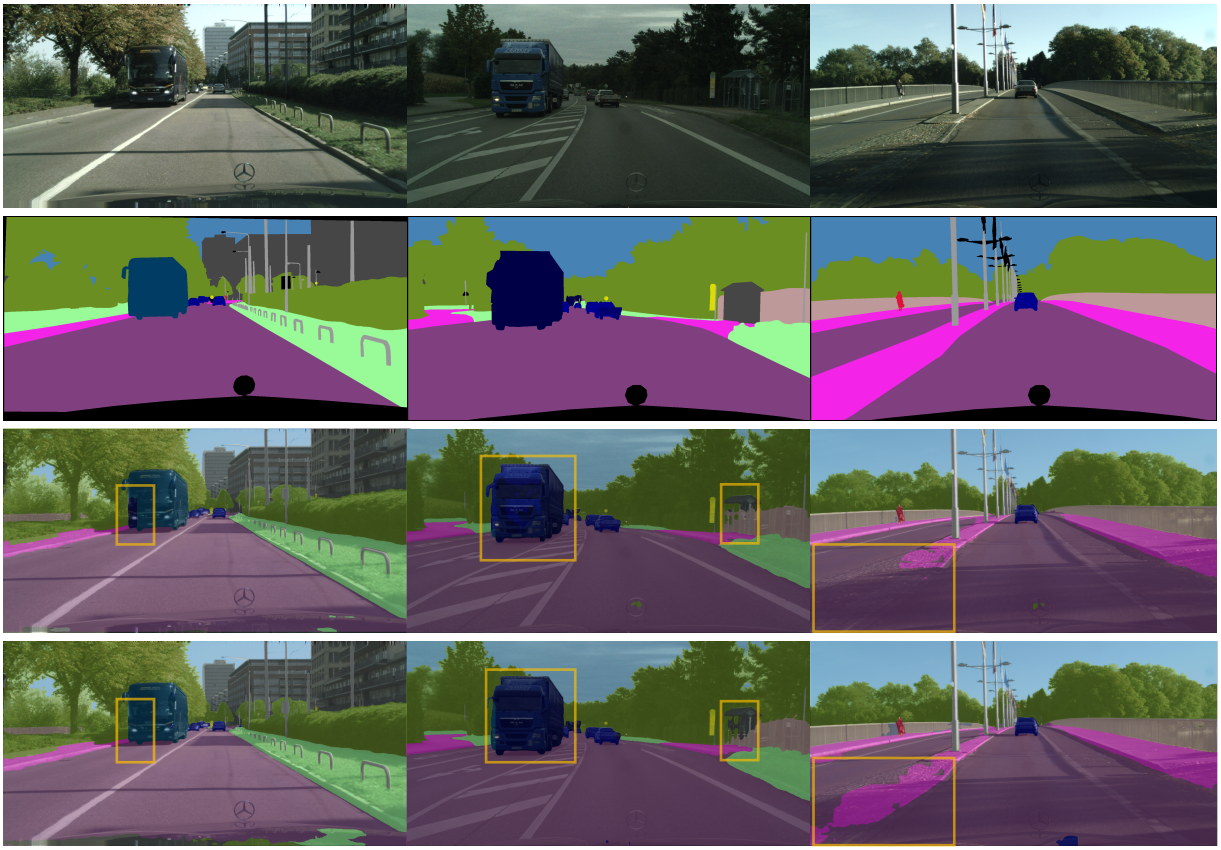

Figure 4. Visualized comparison of segmentation results for PSPNet on Cityscapes validation set. Each column is a group of sample from validation set and the top two rows are the raw image with its corresponding groud truth. The bottom two rows represents the results of PSPNet with ResNet50 backbone and ConT-M backbone separately. The yellow bounding box marks the region where the segmentation results presents a considerable variance.

[3] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2, 3

[5] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 8

[7] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 7

[8] 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*, pages 764–773, 2017. 2

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 6

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[11] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[12] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr. Res2net: A new multi-scale backbone archi-

tecture. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158, 2015. 1

[14] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 7

[15] 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*, pages 770–778, 2016. 1, 2, 4, 6

[16] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 1, 2

[17] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019. 2

[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[19] 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*, pages 3464–3473, 2019. 3

[20] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. *arXiv preprint arXiv:1810.12348*, 2018. 2, 3

[21] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1, 2

[23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 2

[24] Y. Jiang, S. Chang, and Z. Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 2021. 3

[25] S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2661–2671, 2019. 2

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1, 2

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 7

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 7

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[30] Z. Liu, S. Luo, W. Li, J. Lu, Y. Wu, C. Li, and L. Yang. Convtransformer: A convolutional transformer network for video frame synthesis. *arXiv preprint arXiv:2011.10185*, 2020. 3

[31] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[32] I. Loshchilov and F. Hutter. Fixing weight decay regularization in adam. 2018. 7

[33] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018. 2, 3

[34] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018. 8

[35] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019. 3

[36] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 7

[37] P. Shaw, J. Uszkoreit, and A. Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 8

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2

[39] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for machine learning*. Mit Press, 2012. 6

[40] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 2

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2

[42] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9627–9636, 2019. 2, 7

[43] A. Toshev and C. Szegedy. Human pose estimation via deep neural networks'. *CVPR.(Columbus, Ohio, 2014)*, pages 1653–1660. 1

[44] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 2, 3, 6, 7

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 1, 3, 4, 5, 6

[46] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 2

[47] 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*, pages 7794–7803, 2018. 2, 3

[48] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2, 3

[49] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 2

[50] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 2

[51] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 7

[52] H. Zhao, J. Jia, and V. Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10076–10085, 2020. 3

[53] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 8

[54] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840*, 2020. 3

[55] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 2

[56] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3