Multi-scale Spatial Representation Learning via Recursive Hermite Polynomial Networks

Lin (Yuanbo) Wu[1] , **Deyin Liu**[2] , **Xiaojie Guo**[3] , **Richang Hong**[1] , **Liangchen Liu** [4] , **Rui Zhang**[5]
[1]Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education; School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China.
[2]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei 230039, China.
[3]Tianjin University, China
[4] The University of Melbourne, Victoria 3052, Australia.
[5] www.ruizhang.info
{xiaoxian.wu9188, xj.max.guo,hongrc.hfut}@gmail.com, iedyzzu@outlook.com,
liangchen.liu@unimelb.edu.au, rayteam@yeah.net.

Abstract

Multi-scale representation learning aims to leverage diverse features from different layers of Convolutional Neural Networks (CNNs) for boosting the feature robustness to scale variance. For dense prediction tasks, two key properties should be satisfied: the high spatial variance across convolutional layers, and the sub-scale granularity inside a convolutional layer for fine-grained features. To pursue the two properties, this paper proposes Recursive Hermite Polynomial Networks (RHP-Nets for short). The proposed RHP-Nets consist of two major components: 1) a dilated convolution to maintain the spatial resolution across layers, and 2) a family of Hermite polynomials over a subset of dilated grids, which recursively constructs subscale representations to avoid the artifacts caused by naively applying the dilation convolution. The resultant sub-scale granular features are fused via trainable Hermite coefficients to form the multiresolution representations that can be fed into the next deeper layer, and thus allowing feature interchanging at all levels. Extensive experiments are conducted to demonstrate the efficacy of our design, and reveal its superiority over state-of-the-art alternatives on a variety of image recognition tasks. Besides, introspective studies are provided to further understand the properties of our method.

1 Introduction

Dense prediction tasks, such as object localization [Wang *et al.*, 2020; Yu *et al.*, 2017] and semantic segmentation, typically require to describe scale-varied objects with high spatial variances and fine-grained details. To account for objects that may appear with different sizes, multi-resolution representations through CNNs form the basis of a standard solution, due to the capability of CNNs in representing robust and expressive features in hierarchy. Technically, previous methods in this direction mainly resort to *skip connections* [Lin *et al.*, 2017a; Huang *et al.*, 2017], where features with multi-sized receptive fields are fused at different depths, and/or *multi-branch feature fusion* [Li *et al.*, 2019; Wang *et al.*, 2020; Szegedy *et al.*, 2016], where the input and output channels are alternated by dynamic sampling rates, for alleviating scale variations. However, CNNs are limited in preserving spatial precision, since a series of pooling or striding layers would produce coarse, high-level features for deeper layers of the feature network. Such an architecture can reduce the spatial resolution of the resulting feature maps, which is hardly regained and non-invertible. This regret poses an obstacle to dense prediction tasks from achieving higher performance in complex image understanding. For example, matching two scenic person shots needs to simultaneously consider multiscale matching (due to camera distances), and fine-grained matching with accurate spatial dimensions (*e.g.*, the hairstyle, shoes, and jacket patterns).

As a representative solution, dilated convolution is shown to be effective for tasks that require high-resolution predictions [Wang *et al.*, 2020; Takahashi and Mitsufuji, 2021]. The dilation factors are set to grow exponentially as deep layers are stacked, and thereby the network can cover a larger receptive field. However, applying the dilation convolution incurs the prominent *aliasing problem* [Gong and Poellabauer, 2018; Wang *et al.*, 2018], where the signal over the Nyquist frequency becomes indistinguishable with lower frequency after sampling. The dilation convolution with sub-sampling can give rise to such artifacts in feature maps whose receptive field is smaller than the dilation factor [Gong and Poellabauer, 2018]. This is especially obvious for the fine-grained features with much higher frequency. Thus, an appropriate lowpass filter for anti-artifacts (*e.g.*, a standard convolution filter [Wang *et al.*, 2018; Fuchs *et al.*, 2019]) is needed to boost the accuracy of dense prediction tasks.

In this paper, we develop a novel multi-scale spatial representation learning approach to seek accurate and highly expressive features for dense prediction tasks. We present a network based on dilated convolution coupled with Recursive

Hermite Polynomials (called **RHP-Nets**) to incorporate the spatial variance and sub-scale feature granularity. The family of Hermite polynomials possesses the *recursive property* [Pauwels *et al.*, 1995] of producing sub-scale features without losing the spatial resolution. Specifically, by increasing the dilation factor at deeper layers, the proposed RHP-Nets receive larger receptive fields to maintain the spatial resolution across convolutional layers. To explicit mitigate the artifacts caused by the dilated convolution, we apply a set of Hermite polynomials attending to the subset of dilated grids, and recursively generate sub-scale features with granularity. This operation resembles the low-passing filter on the fine-grained features, and also allows the information interchange amongst representations at all layers. By doing so, the aliasing problem is greatly alleviated, and the proposed paradigm not only preserves the recognizable details but also fuses information from different scales into the same representation in an end-to-end manner.

The contributions of this paper are summarized as follows: 1) We introduce Recursive Hermite Polynomial Networks (RHP-Nets) to learn multi-scale spatial representations, which are demonstrated to be beneficial to dense prediction tasks that require robust and expressive features with high spatial accuracy; 2) Multi-scale networks are designed by respecting the spatial variance across layers and exploring the sub-scale feature granularity within a convolutional layer to address the undesirable artifacts.

2 Related Work

2.1 Multi-Scale Representation Learning

Existing works in this direction can be roughly categorized into skip connection based [Lin *et al.*, 2017a; Huang *et al.*, 2017] and multi-branch feature fusion based [Li *et al.*, 2019; Wang *et al.*, 2020; Szegedy *et al.*, 2016] methods. The skip connection structure exploits the inherent design of CNNs to create short paths between different layers. However, a simple connectivity of multiple convolutional layers might not be the optimal way to increase the expressiveness of representations, as too many layers are required to cover a sufficiently large input, and training the network is difficult. Multi-branch fusion is to explicitly model the inter-channel dependencies between convolutional features. For instance, ScaleNets [Li *et al.*, 2019] generate multi-scale representations by down-sampling the input feature maps at different factors while up-scaling the low-resolution representations to recover the lost resolution. Nonetheless, these methods commonly capture multi-scale features in either channel-wise or weighted summation manners, laying intensive emphasis on architecture engineering. Also, by using a medium sized receptive field, the above methods reduce the spatial resolution of the resulting feature maps.

Dilated convolution has been shown effective in many dense predictions tasks that require high resolutions [Wang *et al.*, 2020; Chen *et al.*, 2018b; Takahashi and Mitsufuji, 2021]. It is able to generalize the regular convolution through expanding the kernels with zero insertion. This operation effectively increases the receptive field, so as to perceive a larger spatial context without introducing additional parameters. For instance, PConv [Wang *et al.*, 2020] exploits multi-scale features by manipulating a group of dilated rates to extract diverse features corresponding to different receptive fields. Atrous Spatial Pyramid Pooling (ASPP) [Chen *et al.*, 2018b] sums parallel spatial results for semantic context pooling. The concept of stacking multiple dilated convolutions (SDC) in parallel and combining each output by concatenation is proposed in [Schuster *et al.*, 2019]. However, we remark that these methods [Wang *et al.*, 2020; Schuster *et al.*, 2019; Takahashi and Mitsufuji, 2021] simply use convolution with different dilation rates to compute the multi-scale feature descriptor, while the fine-grained features could be over-smoothed and become indistinguishable after sub-sampling. This prominent problem is also known as aliasing [Gong and Poellabauer, 2018], which degrades the performance of CNN-based recognition tasks [Wang *et al.*, 2018; Fuchs *et al.*, 2019]. Inspired by an interesting property of *recursivity principle* in scale-space [Pauwels *et al.*, 1995], this paper presents recursive Hermite polynomials to prevent the occurrence of aliasing in fine-grained features. It implies that the increasingly blurrier version of an image can be *generated* from the intermediate levels of its scaled variant with less frequency reduced. Hence, we formulate a family of Hermite polynomials to recursively produce sub-scale features within a convolutional layer.

2.2 Multi-Resolution Modeling

Fusing feature maps in different resolutions from early layers are important to dense prediction tasks [Long *et al.*, 2015; Newell *et al.*, 2016]. For example, in Hourglass [Newell *et al.*, 2016], early down-sampled features are first up-sampled and combined via skip connections. Another method for combining feature maps [Sun *et al.*, 2019] attempts to use stage-wise aggregation, *i.e.*, in each stage, feature maps in different resolutions are processed by CNNs individually and then aggregated by a cross-resolution matching with up/down-sampling at the end of each stage. However, this stage-wise aggregation only fuses feature maps globally without local feature fusion. In stark contrast, we produce sub-scale features of different resolutions within a convolutional layer, which are aggregated via trainable Hermite coefficients. This allows us to fuse feature maps with multiple resolutions at all layers. The multi-dilated convolution [Takahashi and Mitsufuji, 2021] can be considered as a multi-resolution modelling where the dense-connected convolutions operate different resolutions using dilated factors depending on skip connections [Huang *et al.*, 2017]. However, such a method can generate undesirable artifacts in feature maps.

3 Method

In this section, we detail how the proposed Recursive Hermite Polynomial Networks (RHP-Nets) learn multi-scale spatial representations. To preserve spatial resolution across deep layers of CNNs, we use a set of dilated filters with larger receptive fields to be applied on input features. To enhance fine-grained features for dense prediction, we address the artifacts caused by dilated convolutions through a sequence of Hermite polynomials to recursively produce sub-scale feature

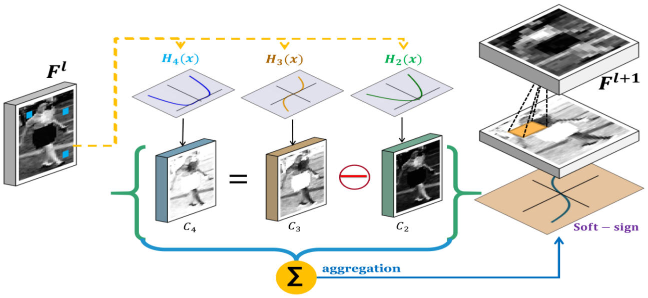

Figure 1: The architecture of a single layer in RHP-Nets. Our contribution is the Hermite Polynomial Block (HPB) that extracts contextual features under a subset of dilated grids, and recursively low-passes them to form the sub-scale output. These sub-scaled responses are aggregated via Hermite coefficients (*i.e.*, C_n), activated by a soft-sign function and fed into the next layer.

maps. These feature maps are aggregated via trainable Hermite coefficients to interchange local representations at all levels. In what follows, we describe the dilated convolution to preserve spatial resolution during the course of convolutions. Then, we describe how the Hermite polynomials are utilized to generate sub-scaled granular features without artifacts.

3.1 Dilated Convolution

The alternatives to obtain a larger receptive field could be larger kernels and dilated convolutions [Yu *et al.*, 2017; Li *et al.*, 2018]. The drawback of using a larger kernel is the high computational cost in terms of the increased number of parameters. To circumvent this problem, dilated convolution advises an enlarged sampling interval with sparsity. In this spirit, we can create the network with dilated convolutional layers, each of which suggests an increased dilated rate to enlarge the receptive field in deeper layers.

To be self-contained, we first define the feature block at a convolutional layer l by denoting $F^l \in R^{C\times H\times W}$, where C is the number of channels, H and W are the height and width. Further denote the convolutional filter as $\mathcal{G} \in R^{C_{ot}\times C_{in}\times K\times K}$, where a set of C_{ot} filters with size $K \times K$ convolve with the input feature, and each filter applies C_{in} kernels to match those input channels. The convolutional operation takes the form:

$$F^{l+1}_{c,x,y} = \sum_{k=1}^{C_{in}} \sum_{i=-m}^{m} \sum_{j=-m}^{m} \mathcal{G}_{c,k,i,j} F^{l}_{k,x+i,y+j}. \qquad (1)$$

where $m = \frac{K-1}{2}$. $F^{l+1}_{c,x,y} \in R^{C_{ot}\times H\times W}$ is the output feature in layer $l+1$, $c = 1, \ldots, C_{ot}$ indexes the output channel, x and y are indices of spatial positions in a feature map.

It is argued that convolution with a dilation rate d and stride d amounts to convolving the sub-sampled input by a factor d without dilation [Chen *et al.*, 2018b; Schuster *et al.*, 2019; Li *et al.*, 2018]. Thus, dilated convolution with striding-free produces a d-downsampled response at full spatial coverage. This critical observation is heavily used by our network design where we employ the output of convolutions with different dilation rates to produce a multi-scale response with spatial resolution preserved. Specifically, given a dilated rate d, the dilated convolution becomes

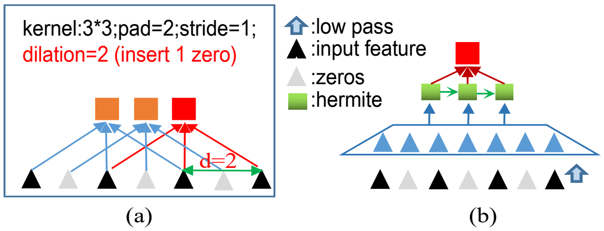

Figure 2: (a) Dilation convolution can cause artifacts by inserting zeros into input features. (b) Hermite polynomials use low-passing and recursively produce sub-scale features that can be integrated into the next layer.

$$F^{l+1}_{c,x,y} = \sum_{k=1}^{C_{in}} \sum_{i=-m}^{m} \sum_{j=-m}^{m} \mathcal{G}_{c,k,i,j} F^{l}_{k,x+\mathrm{d}i,y+\mathrm{d}j}. \qquad (2)$$

Apparently, if we define the receptive field of one element in layer F^{l+1} as a subset elements in the preceding layer F^l, then it drives the receptive field of each element in F^{l+1} with $(2^{l+2}-1)\times(2^{l+2}-1)$. However, the standard dilation convolution may introduce the artifacts caused by the aliasing effect (i.e., insert zeros into input features. Fig.2 (a)). This would hinder the learning of fine-grained features in dense prediction. For this, we propose to apply Hermite polynomials to address the artifacts, and enable dense interchange amongst local representations in each convolution (see Fig.2 (b)).

3.2 Learning Multi-Scale Spatial Features via Recursive Hermite Polynomials

In dense prediction tasks, many methods preserve the spatial resolution across convolutional layers by cascading dilated convolutional layers with different dilation factors. However, the standard dilated convolution can cause the artifacts which impedes the fine-grained features for boosting the prediction accuracy. To alleviate the artifacts, we apply a sequence of recursive Hermite polynomials to low-pass the high frequency and generate the sub-scale granularity.

Hermite Polynomials Hermite polynomials define a class of orthogonal polynomials that can be applied in signal processing [Rasiah *et al.*, 1997], local image matching [Martens, 2006] and polynomial activations in neural networks [Lokhande *et al.*, 2020; Ge *et al.*, 2018]. Mathematically, the normalized Hermite polynomials can be expressed as

$$h_n(x) = \frac{(-1)^n}{\sqrt{n!}} e^{x^2} \frac{d^n}{dx^n} e^{-x^2}, n > 0; h_0(x) = 1. \qquad (3)$$

An important property regarding Hermite transformation is the *recursive* relationship that can be derived from Eq.(3) as

$$h_{n+1}(x) = 2xh_n(x) - 2nh_{n-1}(x). \qquad (4)$$

We take advantage of this recursive property to resemble the low-passing filter on fine-grained features so as to produce sub-scale granularity without artifacts.

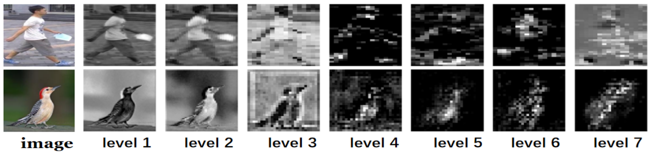

Figure 3: The activation maps at different levels of RHP-Nets. Images are from PRW dataset and CUB dataset. The Hermite polynomial block is applied from level 4 and onward.

Recursive Hermite Polynomial Networks: RHP-Nets As shown in Fig.1, the core of RHP-Nets is a Hermite polynomial block (HPB). In a single convolutional layer, HPB attends to dilated features, and forwards to polynomial activations to low-pass these features. The nature of recursivity in Hermite polynomials guarantees the reduction on spatial frequency over fine-grained details to combat the aliasing caused by cross-layer dilation convolutions. Then, a combination of resultant sub-scale granular features are fused via a set of trainable Hermite coefficients. Each coefficient corresponds to one order of Hermite polynomials. Mathematically, the proposed HPB operation is expressed as

$$F^{l+1}_{c,x,y} = \sum_{k=1}^{C_{in}} \sum_{i=-m}^{m} \sum_{j=-m}^{m} \sigma[\sum_{n=0}^{N} C_n H_n(\mathcal{G}_{c,k,i,j} F^{l}_{k,x+\mathbf{d}i,y+\mathbf{d}j}))], \quad (5)$$

where $H_n(\cdot)$ is computed using Eq. (3). C_n stands for the Hermite coefficient corresponding to the order-n polynomial H_n, $(n = 0, \ldots, N)$. $\sigma(a) = \frac{a}{1+|a|}$ represents the soft-sign function, which plays the role of suppressing the unbounded activations associated with higher-order polynomials [Glorot *et al.*, 2011]. Intuitively, a set of Hermite polynomial transformations amounts to steerable filters for extracting fine-grained features and low-pass them during the scaling. The recursive relationship between Hermite polynomials can reduce the high spatial frequency as the order of polynomials increases. After the removal of high frequent features, we propose an aggregation scheme to fuse these sub-scale representations. We cast C_n to be trainable parameters such that features at different orders are pooled for multi-scale representations. It also ensures that a set of Hermite polynomials are recursive on spatial dimensions without losing the spatial accuracy. Finally, the entire sub-scale features with granularity are aggregated and fed into the next layer .

3.3 Relationship to Existing Methods

At first glance, our method looks similar to PConv [Wang *et al.*, 2020] and SDC [Schuster *et al.*, 2019] in the sense of using dilation to preserve the spatial resolution. However, our method shows clear distinctions in two folds. First, PConv [Wang *et al.*, 2020] and SDC [Schuster *et al.*, 2019] utilize a sequence of dilated convolutions to produce a multi-scale response, which can incur the artifacts caused by the aliasing problem [Gong and Poellabauer, 2018]. Unfortunately, both PConv [Wang *et al.*, 2020] and SDC [Schuster *et al.*, 2019] cannot solve this issue. To circumvent this, in this paper we employ the Hermite polynomials to *recursively* produce sub-scale features within each convolution, during which the same spatially contextual features are sub-sampled to produce the sub-scale output without artifacts. Second, they stack convolutional outputs from each branch, which has limitation in fusing local features. Instead, we aggregate the sub-scale granular features via trainable Hermite coefficients, which allows every local information interchange at all layers.

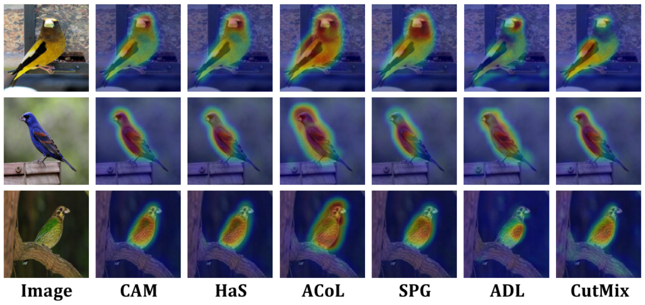

Figure 4: The localization results of RHP-Nets on the CUB dataset.

3.4 Instantiation

The proposed **RHP-Nets** have 7 computational levels in total, denoted as $L_l, l = 1, \ldots, 7$. The network has residual connection analogous to ResNet [He *et al.*, 2016] as the basic module of Conv-BN-ReLU. Each image is resized to be $224 \times 224 \times 3$, and the output is $112 \times 112 \times 32$ after level 1. We remove the max-pooling by replacing it by convolutional filters. The max-pooling operation leads to high-frequency activations that can be propagated to higher layers [Yu *et al.*, 2017]. The max-pooling is only applied in the last level to produce the output of size 14×14. To improve computational efficiency, we adopt a transition layer between the last convolution and the fully-connected layer. The transition layer is composed of a 1×1 convolution followed by a 2×2 average pooling. Finally, a fully-connected layer is added, followed by a classification task using the soft-max function. The HPB is plugged into the network, e.g., ResNet [He *et al.*, 2016] from level 4 to level 7 nested into the 3×3 convolution, where the dilation rates are set $d = \{2, 4, 2, 1\}$, from L_4 to L_7, respectively. In this way, we achieve the spatial accuracy across convolutions to benefit dense prediction.

To produce sub-scale granular features without artifacts, we realize the spatial frequency inside each layer by compositing Hermite polynomials into the dilated grids. This leads to the proposed RHP-Nets by simultaneously achieving the multi-scale representations with high spatial variance and the sub-scale features for boosting dense prediction. The activations on each level are shown in Fig.3. Note that batch normalization is applied before the Hermite polynomial transformation. We tune the choice of the number of Hermite polynomials N by $N \in \{0, 2, 4, 6, 8\}$ in our ablation study. In all experiments, we set $N = 4$ as default.

4 Experiments

4.1 Object Localization

Dataset and Evaluation Metrics We use the Caltech-UCSD Birds-200-2011 (CUB) [Wah *et al.*, 2010] as the benchmark dataset. CUB consists of 200 classes with 5,994 training and 5,794 testing images. We adopt the MaxBoxAcc as the evaluation metric to measure the accuracy between the predicted box and the ground-truth box [Choe *et al.*, 2020]. Given the ground truth box B, the box accuracy at score map threshold τ and IoU threshold δ, $\mathsf{BoxAcc}(\tau, \delta)$ is defined as: $\mathsf{BoxAcc}(\tau, \delta) = \frac{1}{M}\sum_m 1_{IoU(box(s(\mathbf{X}^{(m)}),\tau),B^{(m)})\geq\delta}$, where $\mathrm{box}(s(\mathbf{X}^{(m)}), \tau))$ is the tightest box around the connected component of the mask $\{(i,j)|s(X_{ij}^{(m)}) \geq \tau\}$, $s(\cdot)$ is a scoring function thresholding it at τ, and $X_{ij}^{(m)}$ denotes a pixel in an image $\mathbf{X} \in R^{H\times W}$. Then, the maximal box accuracy is $\mathsf{MaxBoxAcc}(\delta) := \max_\tau \mathsf{BoxAcc}(\tau, \delta)$. We also adopt the top-1 classification/localization as evaluation metrics.

Implementation Details We train our method in fully-supervised setting where each image has full supervision either in a bounding box or binary mask. We consider a foreground saliency mask predictor [Liu *et al.*, 2010], where each pixel is trained with the binary cross-entropy loss against the target mask. The mask is built by labeling pixels inside the ground truth boxes as foreground. To demonstrate the compatibility of our method to different score functions, we set our network as an architecture for classification, and combine with different localization methods: CAM [Zhou *et al.*, 2016], SPG [Zhang *et al.*, 2018b], HaS [Singh and Lee, 2017], ACoL [Zhang *et al.*, 2018a], ADL [Choe and Shim, 2019] and CutMix [Yun *et al.*, 2019]. MaxBoxAcc measures the performance at a fixed IoU threshold $\delta = 0.5$.

Experimental Results Our localization results are shown in Fig. 4. We compare our approach with two recent multi-resolution based methods for dense prediction tasks, i.e., Inception-4v [Szegedy *et al.*, 2016] and SDC [Schuster *et al.*, 2019]. The results in Table 4.1 show that our method achieves consistently performance gain on three measures in the combination with six object localization methods. Though the multi-branch method SDC [Schuster *et al.*, 2019] shows its effectiveness in combining with HaS [Singh and Lee, 2017] and ACoL [Zhang *et al.*, 2018a], it has ignorance to fine-grained features which should be explored to boost the pixel-wise dense prediction on objects.

4.2 Semantic Segmentation

Dataset and Evaluation Metric We use the Cityscapes dataset [Cordts *et al.*, 2016], which contains 5,000 images recorded from street scenes in 50 different cities. Each image has high quality pixel-level annotations. The dataset is annotated with 30 categories, and 19 categories are used for training and evaluation. The training, validation and test set contains 2975, 500, and 1525 images, respectively. In our experiment, the mean of class-wise intersection over union (mIoU) is reported.

Implementation Details The proposed HPB can be integrated with a CNN by simply injecting the polynomial block into the series of convolutional layers. To highlight the effect of Hermite orders, we combine the DenseNet [Huang *et al.*, 2017] with HPB (called D-HPB), i.e., each resolution recursively passes the Hermite polynomials. Data augmentation is performed with random horizontal flipping, cropping and scaling.

Network	Score function	MBA	Top-1 loc	Top-1 cls
Inception	CAM	56.7	40.4	61.8
	HaS	53.4	55.6	70.9
	ACoL	56.2	44.8	56.1
	SPG	55.9	44.9	58.8
	ADL	58.8	39.2	33.1
	CutMix	57.5	48.3	70.2
SDC	CAM	63.0	56.1	58.4
	HaS	64.6	**60.7**	74.5
	ACoL	**66.4**	57.8	64.0
	SPG	60.4	51.5	37.8
	ADL	58.3	41.1	32.7
	CutMix	62.8	54.5	32.7
RHP-Nets	CAM	**67.1**	**57.2**	**69.8**
	HaS	**66.2**	60.5	**76.1**
	ACoL	65.1	**57.9**	**78.0**
	SPG	**63.0**	**60.4**	**78.8**
	ADL	**60.1**	**44.0**	**67.0**
	CutMix	**67.0**	**60.7**	**74.2**

Table 1: Evaluating localization methods on CUB dataset w.r.t three metrics: MaxBoxAcc (MBA), Top-1 localization (loc) and Top-1 classification (cls). Best results are in bold.

Backbone	#param.	mIoU
D-ResNet-50	49.5M	59.7
D-ResNet-101	68.5M	62.4
D3Net-S	9.7M	65.1
D3Net-L	38.7M	68.1
D-HPB (Ours)	37.2M	**70.2**
D-HPB-Light (Ours)	9.2M	**72.3**

Table 2: Segmentation results in complexity and mIoU.

Experimental Results Results in Fig. 5 show that our method achieves superior segmentation performance. Learning to produce sub-scale features as the multi-resolution representations is beneficial to semantic segmentation by leveraging both global and fine-grained information. To thoroughly study the effect of Hermite orders, we consider a variant model that is different in using dilation convolution. The backbones include Dilated-ResNet-50(-101) [Yu *et al.*, 2017] and D3Net-S(-L) [Takahashi and Mitsufuji, 2021]. Experimental results are reported in Table 4.1. D-HPB-Light is a light version by using the channel-reduction mechanism [Huang *et al.*, 2017] to improve the efficiency. Although D3Net [Takahashi and Mitsufuji, 2021] can address the aliasing problem by adjusting the dilation factors, the receptive fields are enlarged repeatedly, leading to an exponentially growing receptive field in almost all layers. In contrast, we design the pluggable polynomial activation as the low-passing filter which is computational efficient.

4.3 Object Detection on MS COCO

The MS COCO dataset [Lin *et al.*, 2014] has 80 categories, which contains 115k images for training (*train2017*), 5k images for validation (*val2017*), and 20k images for testing (*test-dev*). We train the model on *train2017*, and report the re-

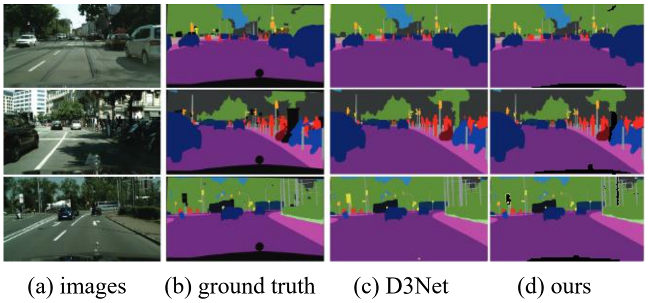

Figure 5: The segmentation results of Cityscapes.

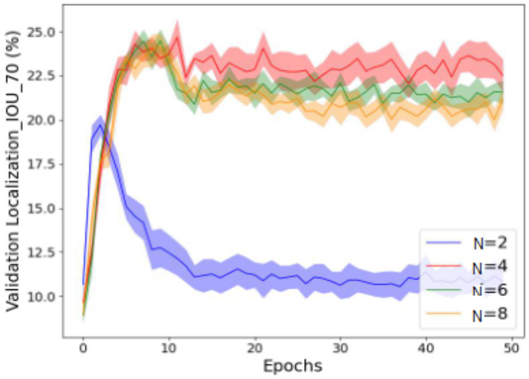

Figure 6: The study on the orders of Hermite polynomials. The localization measure on varied number orders of polynomials show that $N = 4$ is high enough to achieve satisfactory accuracy.

sults on *test-dev*. All reported results follow standard COCO-style Average Precision (AP) metrics.

Implementation Details All experiments are implemented based on mmdetecton [Chen *et al.*, 2018a]. Following [Lin *et al.*, 2017a], the shorter sizes of input images are resized to 800 pixels. For a fair comparison, we re-implement the following detectors equipped with the backbone ResNet-101-FPN [Lin *et al.*, 2017a]: RetinaNet [Lin *et al.*, 2017b], Faster-R-CNN [Ren *et al.*, 2015], Libra-R-CNN [Pang *et al.*, 2019], Grid-R-CNN [Lu *et al.*, 2019], and Mask-R-CNN [He *et al.*, 2017]. Another two backbones with dilation convolution are D-ResNet-50(-101) [Yu *et al.*, 2017] and D3Net-S(-L) [Takahashi and Mitsufuji, 2021]. We also consider a SOTA object detection method, i.e., DeepLabV3 [Chen *et al.*, 2018b] backboned on D-ResNet-50(-101).

Experimental Results Comparison results on MS COCO are shown in Table 4.3. We have the following observations: **1)** Using the proposed RHP-Nets as the backbone, contemporary detectors can achieve superior performance. For example, by replacing FPN with RHP, Mask-R-CNN using ResNet-101 as backbone (denoted as ResNet-101-RHP) achieves 46.1 AP, which is 7.5 point higher than Mask-R-CNN based on ResNet-101-FPN. This clearly shows the importance of the proposed RHP in object detection. **2)** Comparing with the detectors that are backboned on dilated CNNs, i.e., D-ResNet-50(-101) and D3Net-S(-L), our method can still improves the performance. For example, when using Faster-R-CNN based on DenseNet+RHP, our performance is 46.3 AP, whilst Faster-R-CNN with D3Net-S achieves 45.3. **3)** Comparing with DeepLabV3 [Chen *et al.*, 2018b] based on D-ResNet-101 [Yu *et al.*, 2017], our method can effectively address the aliasing problem caused by the atrous convolution, thus improves the object detection performance. For example, Faster-R-CNN [Ren *et al.*, 2015] (backboned on ResNet-101+RHP) improves AP by 0.7 in comparing with DeepLabV3 [Chen *et al.*, 2018b].

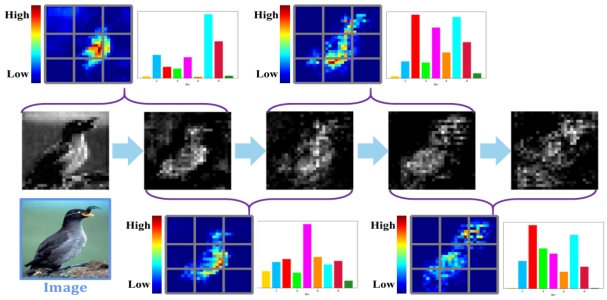

Figure 7: The heatmaps coupled with statistical values in 9 bins show that the adjacent layers of RHP-Nets (from level 3 to level 7) maintain the spatial variance. (The higher values in bins, the stronger spatial correlation). Best viewed in color.

4.4 Multi-scale Person Matching

Dataset and Evaluation Metrics PRW [Zheng *et al.*, 2017] is a person search benchmark. The dataset contains a total of 11,816 video frames and 43,100 person bounding boxes. The training set has 482 different identities from 5,704 raw video frames and the testing set has 2,057 probe IDs along with a gallery repository of 6,122 images. The resolutions of PRW range from 58×21 to 777×574, which is a challenging multi-scale matching problem. Hence, this dataset presents the intrinsic multi-scale challenge. These detected person bounding boxes are determined by employing the Faster-R-CNN [Ren *et al.*, 2015], due to its excellence in detecting varying sized objects in unconstrained scenes. Thereby, it is only necessary to evaluate the performance of our method and state-of-the-arts (SOTAs) on this task. We adopt the widely used protocols: Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP).

Implementation Details All person bounding boxes are resized to 256×128 pixels. To have fair comparison with existing methods ([Lan *et al.*, 2018; Zheng *et al.*, 2017]), we use both annotated and detected boxes to train the identity matching loss. The learning rate is initialized to be 0.001, and the warming up skill is applied by reducing the learning rate as more epochs. Each image is randomly erased with a region to reduce the over-fitting effect.

Experimental Results We compared our method with several SOTA methods in person search: OIM [Xiao *et al.*, 2017], NPSM [Liu *et al.*, 2017], CWS [Zheng *et al.*, 2017], Cross-Level Semantic Alignment (CLSA) [Lan *et al.*, 2018],

Method	Backbone	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
RetinaNet	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
Faster-R-CNN	ResNet-101-FPN	36.3	59.1	42.3	21.8	42.7	50.2
Libra-R-CNN	ResNet-101-FPN	40.3	61.3	43.9	22.9	43.1	51.0
Grid-R-CNN	ResNet-101-FPN	41.5	60.9	44.5	23.3	44.9	53.1
Mask-R-CNN	ResNet-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
DeepLabV3	D-ResNet-50	45.2	63.7	47.7	25.4	46.9	55.8
DeepLabV3	D-ResNet-101	45.4	63.8	48.1	26.0	47.0	55.9
Faster-R-CNN	D3Net-S	45.3	65.0	49.8	28.7	49.2	57.4
Faster-R-CNN	D3Net-L	45.6	65.1	49.8	29.1	49.6	57.3
Faster-R-CNN	ResNet-101+RHP	**46.1**	66.0	51.4	31.7	50.8	58.4
Mask-R-CNN	ResNet-101+RHP	**45.7**	65.2	51.0	31.4	50.6	58.2
Faster-R-CNN	DenseNet+RHP	**46.3**	66.7	51.5	31.4	50.0	57.4
Mask-R-CNN	DenseNet+RHP	**46.1**	65.4	51.4	31.2	50.1	57.3

Table 3: Comparison results with state-of-the-art methods on MS COCO *test-dev*. Best results are in bold.

	Method	Rank-1	Rank-5	Rank-10	mAP
Skip-Connection	OIM	49.9	85.1	89.6	21.3
	NPSM	53.1	85.2	87.7	24.2
	IDE + CWS	45.5	84.1	89.6	18.3
	FPN	53.4	86.4	90.8	31.9
	MuDeep	55.7	84.2	87.9	37.9
	CLSA	65.0	88.7	88.1	38.7
	PPS	72.8	84.6	89.8	48.7
Dilation	PConv	68.2	82.5	85.0	41.1
	CMSNet	70.6	86.7	89.7	42.7
	SDC	73.1	86.7	90.4	46.8
	RHP-Nets	**80.7**	**91.2**	**92.7**	**57.5**

Table 4: Comparison results with SOTA methods (%) for multi-scale person matching on PRW dataset.

MuDeep [Qian *et al.*, 2017], PPS [Shen *et al.*, 2019] and CMSNet [Fan *et al.*, 2020]. Besides, we compare with recent multi-scale learning methods that can be applied into person search: FPN [Lin *et al.*, 2017a] and PConv [Wang *et al.*, 2020]. We use the code released by the authors. Comparison results on rank-1, -5, -10 and mAP values are reported in Table 4.4. It can be seen that our method achieves consistent improvement over STOAs in CMC values and mAP. It also shows that most of skip connection methods align the features to leverage coarse-to-fine details so as to improve the person matching. Though these methods help, they lack explicit mechanisms to fuse diverse information from different scales into the same representation. The family of multi-branch methods, i.e., PConv [Wang *et al.*, 2020] and CMSNet [Fan *et al.*, 2020], can combat this problem by aggregating scale characteristics into one convolutional layer. In stark contrast, our approach constructs spatial polynomials to encode the fine-grained structure. Therefore, multi-scale spatial representations are helpful in multi-scale person search, which requires fine-grained details.

4.5 Ablation Study

In this section, we examine the impact of Hermite polynomial orders and the quality of the learned spatial representations.

The Orders of Hermite Polynomials for Scaling

In this experiment, we empirically determine the number of Hermite polynomials. To observe the effects of using varied orders, we use the measure of IoU on localization validation set of CUB dataset. In Fig. 6, we evaluate the IoU with 70%, 50%, 30% of validation set and calculate the average of the above values. We observe that setting the polynomial order to be $N = 4$ is enough to achieve higher accuracy on localization. Thus, we set $N = 4$ as default in all our experiments.

How Goodness of Learned Spatial Representations?

In this experiment, we validate the quality of spatial representations empirically. To show this quantitatively, we use the spatial relevance to qualify the degree of spatial variance across levels. The spatial relevance [Chiu *et al.*, 2013] is defined as $S(\epsilon_x, \epsilon_y) = \frac{<I(x,y), I(x+\epsilon_x, y+\epsilon_y)>}{<I(x,y), I(x,y)>}$, where we set the offsets ϵ_x and ϵ_y along the x and y axis, respectively. As shown in Fig. 7, the spatial statistics in 9 bins across levels of our RHP-Nets show that we preserve the high spatial resolution entire the network.

5 Conclusion

In this paper, we show the importance of multi-scale spatial representations with granularity in dense prediction tasks and propose a novel architecture called RHP-Nets. We present a network that considers a large context region with high spatial variance across layers, and yet fine-grained feature with granularity within a convolution to aggregate the multiple resolutions. These properties are achieved by applying dilated convolutions across deep layers of CNNs and a family of Hermite polynomials inside a layer to address the artifacts caused by the standard dilation. Extensive experiments in a variety of image recognition tasks confirm the effectiveness of the proposed method.

Acknowledgements

This project was partially funded by NSFC U19A2073, 62002096 and the Co-operative Innovation Project of Colleges in Anhui (GXXT-2019-025).

References

[Chen *et al.*, 2018a] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin.

mmdetection. In *https://github.com/open-mmlab/mmdetection*, pages –, 2018.

[Chen *et al.*, 2018b] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Patt. Anal. Mach. Intel.*, 40(4):834–848, April 2018.

[Chiu *et al.*, 2013] C. Chiu, M. A. Digman, and E. Gratton. Cell matrix remodeling ability shown by image spatial correlation. *Journal of Biophysics*, 2013(doi: 10.1155/2013/532030):1–8, July 2013.

[Choe and Shim, 2019] J. Choe and H. Shim. Attention-based dropout layer for weakly supervised object localization. In *CVPR*, pages 2219–2228, 2019.

[Choe *et al.*, 2020] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim. Evaluating weakly supervised object localization methods right. In *CVPR*, pages 3133–3142, 2020.

[Cordts *et al.*, 2016] 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 *CVPR*, 2016.

[Fan *et al.*, 2020] B. Fan, L. Wang, R. Zhang, Z. Guo, Y. Zhao, R. Li, and W. Gong. Contextual multi-scale feature learning for person re-identification. In *ACM Multimedia*, 2020.

[Fuchs *et al.*, 2019] A. Fuchs, R. Priewald, and F. Pernkopf. Recurrent dilated densenets for a time-series segmentation task. In *International Conference on Machine Learning and Applications*, pages 75–80, 2019.

[Ge *et al.*, 2018] R. Ge, J. D. Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design. In *ICLR*, pages –, 2018.

[Glorot *et al.*, 2011] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *International conference on artificial intelligence and statistics*, pages 315–323, 2011.

[Gong and Poellabauer, 2018] Y. Gong and C. Poellabauer. Impact of aliasing on deep cnn-based end-to-end acoustic models. In *International Speech*, 2018.

[He *et al.*, 2016] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages –, 2017.

[Huang *et al.*, 2017] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

[Lan *et al.*, 2018] X. Lan, X. Zhu, and S. Gong. Person search by multi-scale matching. In *ECCV*, pages –, 2018.

[Li *et al.*, 2018] Y. Li, X. Zhang, and D. Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, 2018.

[Li *et al.*, 2019] Y. Li, Z. Kuang, Y. Chen, and W. Zhang. Data-driven neuron allocation for scale aggregation networks. In *CVPR*, pages 11526–11534, 2019.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages –, 2014.

[Lin *et al.*, 2017a] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[Lin *et al.*, 2017b] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages –, 2017.

[Liu *et al.*, 2010] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Trans. Pat. Anal. Mach. Intel.*, 33(2):353–367, April 2010.

[Liu *et al.*, 2017] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan. Neural person search machines. In *ICCV*, pages 493–501, 2017.

[Lokhande *et al.*, 2020] V. Suresh Lokhande, S. Tasneeyapant, A. Venkatesh, S. N. Ravi, and V. Singh. Generating accurate pseudo-labels in semi-supervised learning and avoiding overconfident predictions via hermite polynomial activations. In *CVPR*, pages 11435–11443, 2020.

[Long *et al.*, 2015] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[Lu *et al.*, 2019] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid r-cnn. In *CVPR*, pages –, 2019.

[Martens, 2006] J.-B. Martens. The hermite transform: A survey. *EURASIP Journal on Advances in Signal Processing*, 2006:1–20, 2006.

[Newell *et al.*, 2016] A. Newell, K. Yang, and J. Deng. Stacked hourglass network for human pose estimation. In *ECCV*, 2016.

[Pang *et al.*, 2019] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *CVPR*, pages –, 2019.

[Pauwels *et al.*, 1995] E. J. Pauwels, L. J. Van Gool, P. Fiddelaers, and T. Moons. An extended class of scale-invariant and recursive scale space filters. *IEEE Trans. Patt. Anal. Mach. Intel.*, 17(4):691–701, 1995.

[Qian *et al.*, 2017] X. Qian, Y. Fu, Y. Jiang, T. Xiang, and X. Xue. Multi-scale deep learning architectures for person re-identification. pages 5399–5408, 2017.

[Rasiah *et al.*, 1997] A.I. Rasiah, R. Togneri, and Y. Attikiouzel. Modelling 1-d signals using hermite basis functions. *IEEE proceedings-Vision, Image and Signal Processing*, 144(6):345–354, 1997.

[Ren *et al.*, 2015] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages –, 2015.

[Schuster *et al.*, 2019] R. Schuster, O. Wasenmuller, C. Unger, and D. Stricker. Sdc–stacked dilated convolution: A unified descriptor network for dense matching tasks. In *CVPR*, pages 2556–2565, 2019.

[Shen *et al.*, 2019] Y. Shen, R. Ji, X. Hong, F. Zheng, X. Guo, Y. Wu, and F. Huang. A part power set model for scale-free person retrieval. In *IJCAI*, pages 3397–3403, 2019.

[Singh and Lee, 2017] K. K. Singh and Y. J. Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017.

[Sun *et al.*, 2019] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[Szegedy *et al.*, 2016] C. Szegedy, S. Ioffe, V. Vanhouck, and A. A. Alemi. Inception-4v, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2016.

[Takahashi and Mitsufuji, 2021] N. Takahashi and Y. Mitsufuji. Densely connected multidilated convolutional networks for dense prediction tasks. In *CVPR*, pages 993–1002, 2021.

[Wah *et al.*, 2010] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds 200. In *Technical report CNS-TR-2010-001, California Insititute of TEchnology*, 2010.

[Wang *et al.*, 2018] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. In *WACV*, 2018.

[Wang *et al.*, 2020] X. Wang, S. Zhang, Z. Yu, L. Feng, and W. Zhang. Scale-equalizing pyramid convolutions for object detection. In *CVPR*, pages 13359–13368, 2020.

[Xiao *et al.*, 2017] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, pages 3376–3385, 2017.

[Yu *et al.*, 2017] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *CVPR*, pages 472–480, 2017.

[Yun *et al.*, 2019] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[Zhang *et al.*, 2018a] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, pages 1325–1334, 2018.

[Zhang *et al.*, 2018b] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang. Self-produced guidance for weakly-supervised object localization. In *ECCV*, 2018.

[Zheng *et al.*, 2017] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *CVPR*, pages 1367–1376, 2017.

[Zhou *et al.*, 2016] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.