# ZoomNet: Deep Aggregation Learning for High-Performance Small Pedestrian Detection

**Chong Shang** SHANG-C13@MAILS.TSINGHUA.EDU.CN
**Haizhou Ai** AHZ@MAIL.TSINGHUA.EDU.CN
**Zijie Zhuang** ZHUANGZJ15@MAILS.TSINGHUA.EDU.CN
**Long Chen** L-CHEN16@MAILS.TSINGHUA.EDU.CN
*Computer Science and Technology Department, Tsinghua University, Beijing, China*

**Junliang Xing** JLXING@NLPR.IA.AC.CN
*Institute of Automation, Chinese Academy of Sciences, Beijing, China*

**Editors:** Jun Zhu and Ichiro Takeuchi

## Abstract

It remains very challenging for a single deep model to detect pedestrians of different sizes appears in an image. One typical remedy for the small pedestrian detection is to up-sample the input and pass it to the network multiple times. Unfortunately this strategy not only exponentially increases the computational cost but also probably impairs the model effectiveness. In this work, we present a deep architecture, refereed to as ZoomNet, which performs small pedestrian detection by deep aggregation learning without up-sampling the input. ZoomNet learns and aggregates deep feature representations at multiple levels and retains the spatial information of the pedestrian from different scales. ZoomNet also learns to cultivate the feature representations from the classification task to the detection task and obtains further performance improvements. Extensive experimental results demonstrate the state-of-the-art performance of ZoomNet. The source code of this work will be made public available to facilitate further studies on this problem.

**Keywords:** pedestrian detection, small object detection, deep learning

## 1. Introduction

Pedestrian detection is a longstanding topic in computer vision, which acts as an important component in many practical systems, *e.g.*, video surveillance, autonomous driving, and robotics. Due to the fast development of deep convolutional neural networks (CNNs), the performance of pedestrian detection in real-world scenarios has been significantly improved in the past decades (Viola and Jones, 2001; Wu and Nevatia, 2007; Dalal and Triggs, 2005; Felzenszwalb et al., 2008; Duan et al., 2012; Xing et al., 2010; Marin et al., 2013; Yan et al., 2013; Dollár et al., 2014; Zhang et al., 2015; Li et al., 2015; Cai et al., 2016; Wang et al., 2017; Zhang et al., 2018a; Bhattacharyya et al., 2018). However, there still remains notable gaps in both detection accuracy and speed for state-of-the-art pedestrian detection models when deployed into practical applications, such as autonomous driving, video surveillance and traffic monitoring, in which the model needs to detect pedestrians of varying sizes.

Recently, Convolutional Neural Networks (CNNs) have shown great potential for handling changes in illumination, backgrounds and deformation in many visual tasks. Several works, such as Girshick et al. (2014); Ren et al. (2015); Redmon et al. (2016); Liu et al.

(2016); Kong et al. (2016); Dai et al. (2016), presented CNN detectors, which followed the same pipeline that transfer ImageNet pre-trained classification models into detection. However, these models might loss spatial information due to a sequence of down-sampling operations, and detecting pedestrians of various sizes thus posed great challenges to these models. In many practical applications, such as autonomous driving, the system has to simultaneously detect very small targets, e.g. 50 or even fewer pixels in height. The reasons that small object detection is difficult include: (1) little visual evidence where small objects with few pixels lack of many distinctive details in their appearances; (2) greater stride in existing detectors (Ren et al., 2015; Liu et al., 2016; Redmon et al., 2016), which down-sample the feature maps to a great extent to obtain stable representations, and may lose even more details for small objects; and (3) unmatched backbones of existing detectors that are usually pre-trained on large objects, *e.g.* $224 \times 224$ for ImageNet, making them inherently biased to objects with larger scales and may leave out small ones.

A straightforward way to handle this problem is to up-sample the input so that small objects become larger in appearance (Yan et al., 2013; Li et al., 2017; Zhang et al., 2017). However, it exponentially increases the computational cost thus cannot be adopted in real-time systems. Another more efficient way to retain more information for small objects is to adopt a finer stride of features by ether removing the last down-sample layer in a CNN (Dai et al., 2016) or constructing a pyramid of features/images (Liu et al., 2016; Lin et al., 2016; Fu et al., 2017; He et al., 2017; Liu et al., 2018; Singh and Davis, 2017; Cao et al., 2018). Some other efforts (Zhang et al., 2018b) attempted to adopt reinforcement learning to search for the activated neurons to handle small pedestrian, which required a sequence of explorations before finding the suitable neurons, making it very time-consuming.

In this work, we focus on the problem of small pedestrian detection. We first study the advantages and limitations of feature pyramid networks (FPN) (Lin et al., 2016), which adopted multi-level feature maps to enhance its capacity for objects of varying sizes. Unfortunately, we find it is still inadequate for smaller objects due to the fact that much information is already lost at the beginning layers. Besides, FPN adopts a large object oriented backbone network, which is pre-trained on ImageNet, and overlooks the object scale differences between classification and detection tasks. To make it worse, the beginning layers are usually fixed during training which may lead to a sub-optimal result for small objects.

To alleviate these shortcomings, we propose a novel deep architecture which aggregate the coarse pyramid features with finer stride ones. The original pyramid contains much high level information but fewer details while the finer features retain more details with an additional path. Then the final features are enhanced by aggregating them. In particular, the finer stride features are obtained via an independent complementary path from the input to the feature pyramid. The down-sampling layer in this path is removed in order to maximize the spatial information flow. More importantly, separating this branch from the backbone allows the network to explore new representations and ease the optimization, thus leads to a better result.

We summarize the main contributions of this work as follows.

- We study the limitations of widely-used CNN detectors, and point out the unmatched domain transferring between general object classification and detection tasks. For

the latter task, one need to handle more scale variant, especially very small objects, resulting the CNN detectors unsuitable for handling small targets.

- We propose an aggregation learning architecture which improves the information flow by aggregating an additional path with the backbone network for small pedestrian detection. The extra path compensates small objects for spatial information loss.
- We propose a light-weighted CNN detector, referred as ZoomNet, to achieve the state-of-the-art result on several challenging pedestrian detection benchmarks with a little extra computational cost.
- We study alternative architectures with the potential of improving small pedestrian representations and provide some discussions on dealing with the drawbacks of pedestrian detectors.

## 2. Related Work

A large number of developments for object detection using different learning models has been published in the last decades. Here we only discuss some of the most related works to ours in this work.

*General CNN detectors.* CNNs extract higher abstractions by stacking multiple transformation matrices, resulting high level semantic representations. Regular CNN detectors (Girshick et al., 2014; Ren et al., 2015; Redmon et al., 2016; Liu et al., 2016) started with pre-trained models on ImageNet, attempted to decode the classes and location for each object from high level features. Following works (Dai et al., 2016; Kong et al., 2017; Lin et al., 2016; He et al., 2017; Liu et al., 2018) further improved detection task with introducing finer stride features. Feature pyramid networks (FPN) (Lin et al., 2016) constructed a pyramid architecture with top-down connections, enriched the spatial information using shallow features. The SNIP training strategy (Singh and Davis, 2017) addressed the fact that CNN detectors were scale specific, thus could not handle extreme scale changes. Their work inspire us to further study how to empower the network to handle extreme small objects. The DetNet (Li et al., 2018) pointed out inherent drawbacks of ImageNet pre-trained models for fine-tuning detectors, and proposed a network with finer feature stride to remedy spatial information loss. However, they ignored the difference of scale distribution between these two tasks and failed to realize the fact that information loss for small objects had already happened at the shallow layers of CNNs.

*Pedestrian detection.* Pedestrian detection is one of most important applications that are applied to real-world problems. Before deep neural networks, features including Haar (Viola and Jones, 2001), Edgelets (Wu and Nevatia, 2007), ACF (Dollár et al., 2014), CheckBoard (Zhang et al., 2015); models including DPM (Felzenszwalb et al., 2008), boosting (Viola and Jones, 2001) and decision trees (Marin et al., 2013) were exploited. Recently, CNN detectors have shown great potential in solving illumination changes, difficult backgrounds, pose variations and scale variations in many general object detection tasks. As for pedestrian detection, existing works, such as Li et al. (2015); Cai et al. (2016) attempted to handle multi-scale problems with different level of features. Zhang et al. (2017) adapted a Faster RCNN detector to achieve a notable performance, inspiring us to further inspect basic

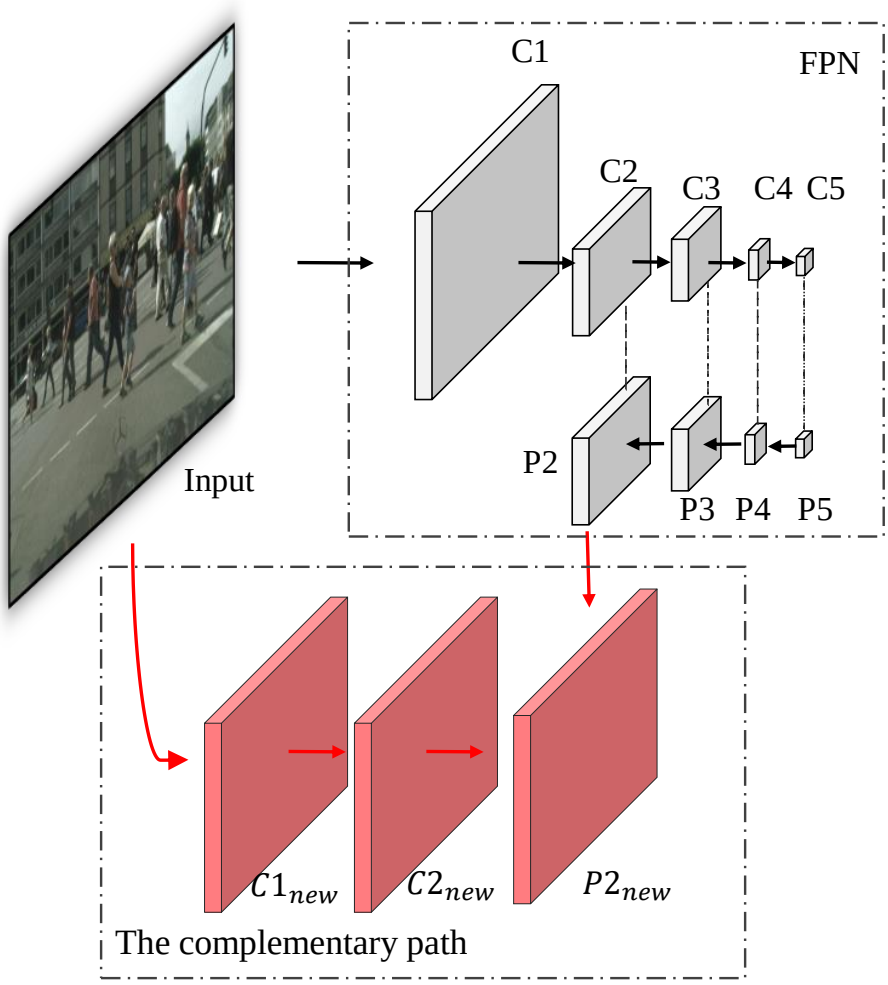



Figure 1: Our proposed network architecture. The red arrows indicate the complementary path, from the input to the pyramid features, which compensates small objects for information loss.

problems in this task. A repulsion loss (Wang et al., 2017) was proposed to guide the network to distinguish occluded persons. More recently, Zhang et al. (2018a) adopted attention mechanics to detect persons which were partially visible.

*Small Object Detection.* One possible and straightforward way for small object detection is to up-sample the input (Yan et al., 2013; Zhang et al., 2017; Wang et al., 2017). However, it exponentially increases the computational cost, restricting its applications in real-time systems. A reinforcement learning strategy (Zhang et al., 2018b) was adopted to search for the activated neurons to handle small pedestrian, which required a sequence of explorations before finding the suitable neurons, making it time-consuming. The FoveaNet (Li et al., 2017) attempted to only up-sample a local region instead of the whole image. It required a strong assumption that the sub-region that needed to up-sample was already known. In parallel with our work, the Multi-Branch and High-Level Semantic network (Cao et al.,

2018) explored to use different branches in order to construct a feature pyramid with the same high-level semantics. To distinguish with theirs, our model shares top-level features across different scales and utilizes only a shallow path to compensate detail loss, resulting a computational efficiency detector.

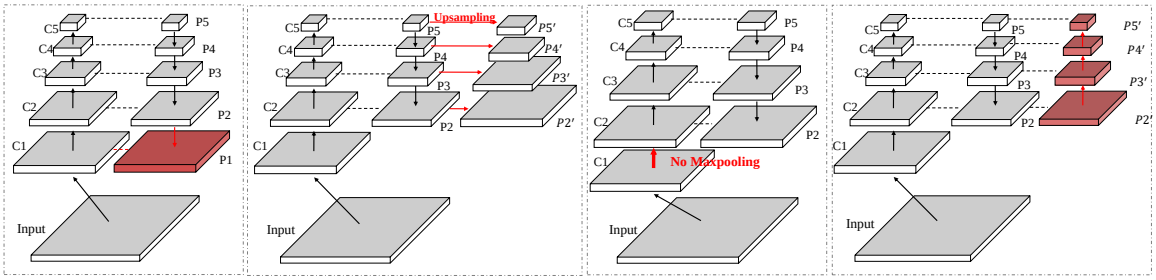
## 3. The ZoomNet Model

The proposed ZoomNet model is illustrated in Figure 1. ZoomNet relieves shortcomings of FPN with aggregation learning for detecting small object. In particular, we introduce a complementary path from the input to the pyramid features to retain more spatial information. Then it is aggregated with the backbone using a top-down connection, making it compatible with the original detector. In contrast to FPN of which the shallow layers are fixed during training, the complementary path is optimized with the entire network at the same time, which allows the model to cultivate feature representations for transferring large-object oriented networks for objects of varying sizes.

### 3.1. A complementary path

We take ResNet50 (He et al., 2016) as the backbone and denote feature stages using $\{C1, C2, C3, C4, C5\}$ with a stride of $\{2, 4, 8, 16, 32\}$. Our model is built on FPNs, which introduce a top-down connection to join different stages ($C2$ – $C5$) to construct a pyramid of deep features ($P2$ – $P5$). Then the classes and positions of objects are generated by these pyramid features proportionate to the scales of objects. To enhance small object detection one method is to adopt finer strides of pyramid features. A straightforward way to expand the FPN is to joining a lower layer (*e.g.*, $C1$) in the pyramid. However, in our controlled experiments, we find out it provides little improvement for both small and large objects, of which one possible explanation is that the details for small objects are already lost in the beginning stage.

Based on this consideration, we propose a novel deep aggregation learning architecture that bridge the gap on small object detection, referred as ZoomNet. As illustrated in Figure 1, ZoomNet contains two paths from the input to pyramid features. The first path is FPN which already forms a strong representation for larger objects. The second path, colored in red, serves as a complementary role on handling small objects. It operates at the shallow layers and outputs a finer and deeper stage of feature $C2_{new}$, which then is enhanced with a top-down connection to form the new pyramid feature $P2_{new}$. More specifically, we design the proposed complementary path to share similar structure as the backbone ($Input$–$C2$) and allows it to inherit structural benefits of ResNet. To distinguish with the backbone, we remove the max-pooling layer after $C1$ and replace the following convolutional layer with a dilated convolutional layer to fill the holes as Yu and Koltun (2015). The resulting feature of the complementary path ($C2_{new}$) has a stride of 2. As for the dilation rate, we set it to 0 and 1, find it has little impact on the final performance. Following FPN, we up-sample $P2$ by a factor of 2, then add it with $C2_{new}$, resulting an enhanced feature level $P2_{new}$. Then we add a region proposal network (RPN) after $\{P2_{new}, P3, P4, P5\}$. Different from existing works, which keep the parameters of shallow layers $(C1, C2)$ fixed during training, we allow the complementary path to update. Note that allowing the shallow layers in the original FPN (baseline) to update at the beginning stage will sabotage the performance by a large




(a) FPN-C1 (b) FPN-Zoom (c) FPN-x2 (d) FPN-PA

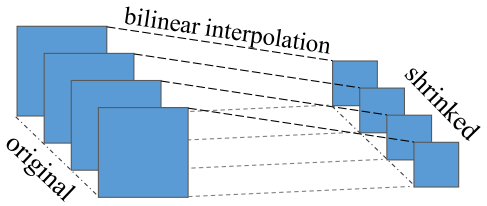
Figure 2: Alternative network architectures that expand pyramid features, (a)–(c), or aggregate an extra path (d). We use red color to indicate adjustments on the FPN detector.

margin. And updating them in the middle of training also generates little performance gain. One explanation is that object of different scales compete with each other, leading to saturation of network capacity. And separating the network into different paths can ease the optimization.

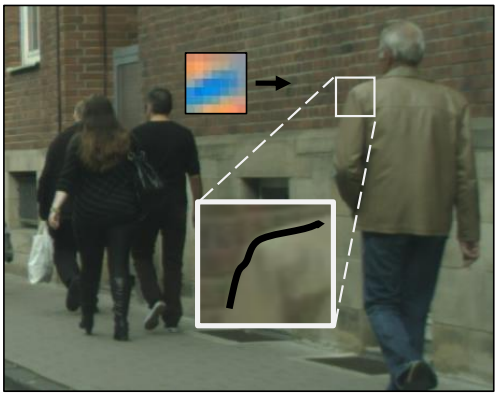
### 3.2. Alternative models

One key factor in this work is to adopt finer stride features. However, not all architecture designs will lead to performance gain. We find that the feature maps not only need to be finer but also need to be deeper. For comparison, we also investigate alternative models that adopt finer features: 1) **FPN-C1**: we join $C1$ in FPN so that the finest pyramid feature is $P1$ with a stride of 2; 2) **FPN-Zoom**: we only up-sample the pyramid features without introducing any paths to validate whether the performance gain is obtained only by adding more anchors; and 3) **FPN-x2**: another architecture is to remove the max-pooling layer after $C1$ so that the spatial dimensions of all pyramid features $P2 - P5$ are doubled. This model leads an improvement for all scales over the baseline with a large margin. However, the computational cost is almost 4 times larger than the baseline, which limits its applications in real-world system. We list the above architectures in Figure 2 (a–c).

Another key factor is to adopt an additional path so that it can be optimized during training time. For comparison, we also perform experiments with the following models to validate the importance of adjusting shallow features for small objects: 1) **FPN-CP-FIX** in which we fix the complementary path in the ZoomNet during training so that it will not be adjusted for detecting small objects; 2) **FPN-AllTrain** in which we also allow the shallow layers to update in the baseline model, which turns out to harm the performance by a large margin; and 3) **FPN-PA** in which we also aggregate a path from middle levels to top levels following Liu et al. (2018), shown in Figure 2 (d). The motivation behind FPN-PA is to shorten the forward propagation path.




Adaptive kernel shrinking



Large objects

Samll objects

Figure 3: Illustration of the adaptive kernel shrinking trick. Since the network is pre-trained on large objects, the kernels in the first convolutional layer is down-sampled to detect small objects.

### 3.3. Adaptive kernel shrinking

Another insight we provide in this work is to shrink the convolutional kernels adaptively in the first layer. Figure 3 illustrates the intuition of kernel shrinking. The first convolutional layer can be viewed as template matching, which detects specific patterns. Since the original kernels are pre-trained for objects with larger scales (*e.g.*, $224 \times 224$ in ImageNet), transferring them to small objects may need adjustments. A common trick is to shrink the size of kernels in the first layer, which is an equivalence of scaling the input (Benenson et al., 2012; Dollár et al., 2014). We define the first convolutional layer with smaller kernel size (*e.g.*, $5 \times 5$), which is initialized by down-sampling the original $7 \times 7$ kernels into $5 \times 5$ shape with bilinear interpolation.

### 3.4. The networks architecture

Given an RoI (Region of Interest) generated by the RPN, in contrary to FPN we first use all the pyramid levels $\{P2_{new}, P3, P4, P5\}$ to crop features into the same spatial dimension. Then they are concatenated along the channel axis followed by a $1 \times 1$ convolution to reduce the channel dimension to the original number. This pooling operations allow an RoI to directly access all feature levels, leading to a stable representation. The pooled

region features are propagated through two $1024 \times 1024$ fully connected layers followed by a classification layer and a box regression layer.

## 4. Experiments

In our experiments we use CityPersons (Zhang et al., 2017) and Caltech (Dollar et al., 2012) to demonstrate the effectiveness of our proposed method on small pedestrian detection. First, we introduce these datasets, evaluation metrics, and implementation details. Then we perform ablation studies on Citypersons to validate the contributing factors that are proposed in this work. Finally, we compare our ZoomNet with the state-of-the-art methods.

### 4.1. Datasets, experimental settings and implementation details

#### 4.1.1. Datasets

Among public datasets, CityPersons is a relatively large benchmark, re-labeled on Cityscapes (Cordts et al., 2016), and consists of 5,000 images and $\sim$ 35,000 persons, more that 30% heavy occlusion cases (visibility $<$ 65%). Besides, it covers a large range of scales, thus making it particularly suitable for studying how CNN detectors react to extreme scale variant.

Caltech pedestrian detection dataset (Dollar et al., 2012) is another widely-used dataset for pedestrian detection, has witnessed the development in this area for decades. It contains 2.5 hours video captured from a moving vehicle. An aligned annotation was provided by Zhang et al. (2016b), consisted of 42,782 frames and 4,024 frames for training and testing respectively. Following this standard setting, we perform experiments with both the new and the original annotations, and report results on the *Reasonable* subset. In contrast to Citypersons, Caltech is dominated by small scale persons, and suffers less from large scale variation problems. The median height of persons is 48. See more statistics in Dollar et al. (2012).

#### 4.1.2. Experimental settings

All of our experiments in this paper are performed on the reasonable train/val splits for training/testing. For fair comparison, the MR (log-average miss rate) between $[10^{-2}, 10^{0}]$ FPPI (false positive per image) is used as evaluation metrics following Dollar et al. (2012); Zhang et al. (2017). We report MRs on different subsets, including the *small* subset ($75 > height > 50$, $0.65 >$ visibility $> 0.20$) and the *reasonable* subset (height $> 50$, visibility $> 0.65$).

#### 4.1.3. Implementation details

We adopt ResNet50 pre-trained on ImageNet as the backbone of our detector. For the RPN, we define anchors of scales {32, 64, 128, 256} that correspond $P2$ to $P5$ respectively, then add more scales of $\{2^0, 2^{1/3}, 2^{2/3}\}$. For aspect-ratios, we simply adopt {1, 2.5} ($height/width$), which is suitable for sitting persons and pedestrians. The resulting heights of anchors span from 32 to 642. Note that our method increase the number of anchors in the lowest pyramid level, thus we adapt our sampling strategy so that anchors in each

Table 1: Comparison with alternative architectures. The performance is measured by Miss Rate (MR, in %). The input size is set to $896 \times 1792$.

| Method | Reasonable | Small | Time (sec. per img) |
|---|---|---|---|
| FPN (baseline) | 15.9 | 26.3 | 0.21 |
| FPN-Zoom | 17.1 | 28.6 | 0.25 |
| FPN-C1 | 16.0 | 26.2 | 0.24 |
| FPN-x2 | **14.8** | 23.5 | 0.52 |
| FPN-PA | 15.6 | 26.4 | 0.23 |
| FPN-AllTrain | 18.4 | 29.1 | 0.21 |
| FPN-CP-FIX | 15.7 | 25.8 | 0.32 |
| ZoomNet (proposed) | **14.9** | **23.1** | 0.32 |

level can be equally sampled when training the RPN. For the rest configurations, we simple follow He et al. (2017). As for the complementary path $Input - C2_{new}$, since it share the same structure with $Input - C2$, we also initialize it using the ImageNet pre-trained model. In contrast, their parameters are allowed to update during training. For experiments on Citypersons, we adopt a initial learning rate of 0.005, then decrease it with a factor of 0.1 at the 8 and 12 epoch. We train the detector for 14 epochs. The resulting detector is stable to the parameter choices. For the complementary path, we initialize it with corresponding backbone parameters, and allow it to update during training time. In the ablation study, the input size is set to $896 \times 1792$ for both training and testing. And for experiments on Caltech, we set the initial learning rate to 0.002, then decrease it by a factor of 0.1 at the 4th and the 6th epochs. The input size is set to $960 \times 720$, which is 1.5 times of the original size. We train models for totally 8 epochs and report the final MRs. The rest configurations are set as the ones in Citypersons unless specified. For data augmentation, we only randomly horizontally flip the image and do not adopt any other kinds of augmentation, such as multiple scale training, in order to reduce the influence of other factors.

### 4.2. Ablation studies

#### 4.2.1. Why an extra path is necessary?

One key insight for object detection is to retain spatial dimension, *i.e.*, using finer stride features. To validate the contributions of adopting finer stride features, we compare the proposed ZoomNet with FPN-C1, FPN-Zoom and FPN-x2. As shown in Table 1, *FPN-C1* involves a finer pyramid features. However, it provides little improvement over the baseline. One possible explanation is that $C1$ is too shallow and contains little useful information. *FPN-Zoom* directly up-samples pyramid features without introducing any additional paths, resulting a doubled spatial dimension. However, it harms the results. The reason might be up-sampling the features does not make the pyramid deeper and provides little additional spatial information. *FPN-x2* removes the maxpooling layer after $C1$ stage, resulting a pyramid twice larger than the baseline. It provides 2.8 MR improvement on the small subset. However, it is time-consuming due to expansion in spatial dimension, which makes it difficult to apply in a real-world system. Compared to FPN-x2, the finest feature of

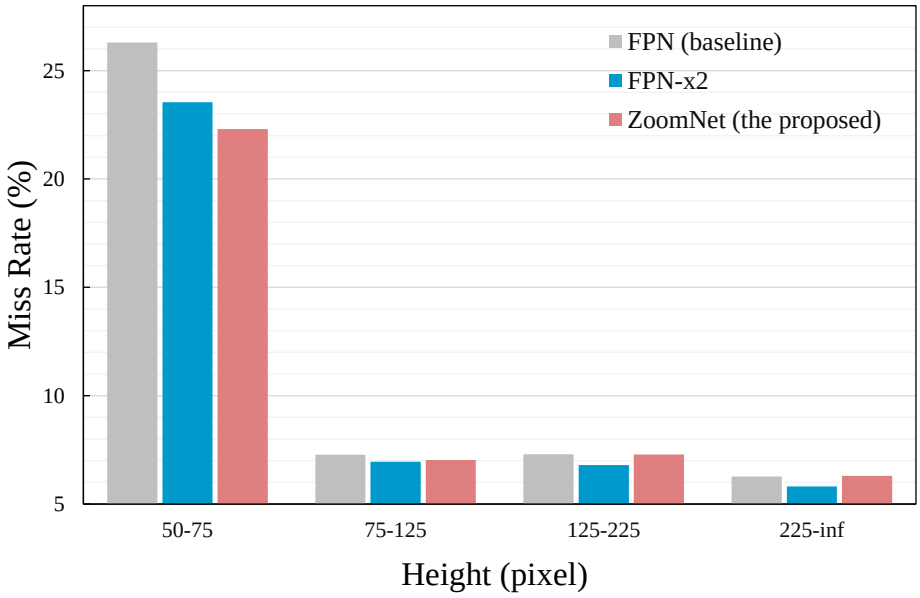



Figure 4: Performance over different scales, in term of miss rate, the lower the better. (Better viewed in color).

ZoomNet not only has the same spatial dimension but also has the same depth. However, FPN-x2 is outperformed by ZoomNet on the small subset with a notable margin.

To further validate the contribution of the complementary path, we compare ZoomNet with FPN-CP-FIX, which shares the same architecture. But the complementary path is fixed during training. FPN-CP-FIX provides only a small improvement over the baseline model, and is outperformed by ZoomNet with a notable margin. It demonstrates that the performance gain is obtained not only by removing the max-pooling layer but also by cultivating the complementary path for small objects. However, allowing the corresponding layers to update in the baseline model, as FPN-AllTrain, harms the results instead of improving them. One possible explanation might be that objects of different scales compete with each other in the baseline model while our aggregation learning framework separates the backbone into two branches, leading the network to obtain a stronger representation. We also perform experiments in which we first fix the shallow layers of FPN in the beginning epochs and then allow them to update in the rest epochs. However, we find that it provides little performance gain, which can support our argument that an extra path is necessary for handling extreme scale variant.

In Figure 4, we show the MRs over different scales. The MR of the baseline model increases dramatically once the heights of objects are lower than 75. FPN-x2 improves the results via retaining more spatial information after C1 stage. However, the improvement of FPN-x2 is at the cost of 4 times computational complexity increasing. Among these models, ZoomNet achieves the best MR on the small subset (height $\in (50, 75)$) with relatively little computational complexity increasing. Besides, MRs for larger objects are also reduced. That is because the separated branches may also ease the optimization for varying scales of

Table 2: Comparison on adaptive kernel shrinking across different kernels in the first convolutional layer (in %). The input size is set to $896 \times 1792$.

| Method | Kernel | Reasonable | Small |
|---|---|---|---|
| FPN (baseline) | – | 16.1 | 26.8 |
| ZoomNet | $7 \times 7$ | 15.2 | 23.3 |
| | $5 \times 5$ | 15.1 | 23.5 |
| | $3 \times 3$ | **14.9** | **23.1** |

Table 3: Comparative results with other methods. Addt. is short for additional supervision. Scale denotes the up-sampling factor with respect to the original image size ($1024 \times 2048$).

| Method | Addt. | Scale | Reasonable | Small |
|---|---|---|---|---|
| Zhang et al. (2017) | – | 1 | 15.4 | 25.6 |
| | Segment. | 1 | 14.8 | 22.6 |
| | – | 1.3 | 12.8 | – |
| Wang et al. (2017) | – | 1 | 14.6 | – |
| | Repulsion | 1 | 13.2 | – |
| FPN (baseline) | | 1 | 14.8 | 22.1 |
| | | 1.3 | 13.3 | 16.7 |
| ZoomNet | | 0.875 | 14.9 | 23.1 |
| | | 1 | **13.1** | **19.1** |
| | | 1.3 | **11.6** | **15.1** |

objects. We can also observe that the improvements are not as obvious as those of FPN-x2 when object heights are greater than 125. The reason is that FPN-x2 increases the spatial dimensions of all pyramid levels while ZoomNet only increases the lowest level.

#### 4.2.2. Adaptive kernel shrinking for small objects

We propose to adjust the first convolutional layer for small objects via adapting the learned kernel size. We compare models with kernel size $7 \times 7$, $5 \times 5$ and $3 \times 3$ in Table 2. Kenerls of shape $5 \times 5$ and $3 \times 3$ are obtained by down-sampling the original $7 \times 7$ kernels using bilinear interpolation. All ZoomNet models outperform the baseline FPN, which demonstrates that shrinking the learned convolutional kenerls will not undermine the network. Among these models ZoomNet-3x3 achieves a slight better result, and also enjoys the benefit of less computational complexity due to smaller kernels.

### 4.3. Comparisons with other methods

#### 4.3.1. Results on Citypersons

We compare our results on Citypersons with the state-of-the-art models in Table 3. Zhang et al. (2017) adopted Faster RCNN detector, and make several adaptations, including re-

Table 4: Comparative results with other methods on Caltech pedestrian detection dataset using new annotations. Scale denotes the up-sampling factor with respect to the original image size ($640 \times 480$). The performance is measured by Miss Rate (MR, in %).

| Method | Scale | Reasonable |
|---|---|---|
| CompACT-Deep (Cai et al., 2015) | – | 9.2 |
| MS-CNN (Cai et al., 2016) | 1.5 | 8.1 |
| RPN-BF (Zhang et al., 2016a) | 1.5 | 7.3 |
| HyperLearner (Mao et al., 2017) | – | 5.5 |
| AdaptedFstrRCNN (Zhang et al., 2017) | 2 | 5.8 |
| RepLoss (Wang et al., 2017) | 2 | 5.0 |
| FPN (baseline) | 1.5 | 4.9 |
| ZoomNet | 1.5 | **4.5** |
| | 2 | **4.3** |

moving last two max pooling layers, using a larger set of anchors and adopting Adam optimizer during training. The baseline model (FPN) already outperforms theirs with less tailored designs. Further our *ZoomNet* outperforms theirs by a 2.2 on the *reasonable* subset and a large margin of 6.3 on the *small* subset. Moreover, the proposed model is even comparable with theirs using a down-sized input (0.875 scale factor) while they achieved the comparable MR with the aid of segmentation result. Wang et al. (2017) also adopted Faster RCNN detector with the ResNet50 backbone, and made some adaptations for pedestrian. And they used a mini-batch of 4 images, four times larger than ours, which might probably bring extra advantages in term of training setups. Besides, instead of improving detection for small objects they focused on handling occlusion with another learning objective. Again, our ZoomNet already achieves an approximate MR of 14.9 with a down-sized input (0.875 scale factor) on the *reasonable* subset. They did not show MR of the *small* subset. On the *reasonable* subset our ZoomNet outperforms theirs with the same input size. Since their goal is to handle occlusion instead of small objects, we can conclude that our ZoomNet also achieves a better result on the *small* subset without any additional supervisions.

#### 4.3.2. Results on Caltech

In Table 4 and 5, we compare our ZoomNet with the state-of-the-art models on Caltech dataset. Our proposed ZoomNet outperforms the state-of-the-art models by a notable margin. Since the scale distribution of Caltech is more concentrated than that of Citypersons, existing models, such as HyperLearner (Mao et al., 2017), AdaptedFstrRCNN (Zhang et al., 2017), and RepLoss (Wang et al., 2017) suffered less from large scale variation; and our baseline model, FPN, already achieves a lower MR. However, the proposed ZoomNet still yields a notable improvement over our strong baseline with 0.4 drop in MR, which demonstrates the effectiveness of the proposed model. Moreover, Zhang et al. (2017) achieved 5.8 MR with an up-sampling factor of 2 while our ZoomNet achieves a 1.1 lower MR with an even smaller input size.

Table 5: Comparative results with other methods on Caltech pedestrian detection dataset using original annotations. The performance is measured by Miss Rate (MR, in %).

| Method | Reasonable |
|---|---|
| CompACT-Deep (Cai et al., 2015) | 11.7 |
| UDN+SS (Ouyang et al., 2018) | 11.5 |
| MS-CNN (Cai et al., 2016) | 9.9 |
| RPN-BF (Zhang et al., 2016a) | 9.7 |
| ADM (Zhang et al., 2018b) | 8.6 |
| FstrRCNN-ATT (Zhang et al., 2018a) | 10.3 |
| FPN (baseline) | 9.4 |
| ZoomNet | **8.5** |

## 5. Conclusion

In this work, we focus on small pedestrian detection, which is critical to many real-world applications. We provide an intriguing finding that information loss for small objects already happens in the beginning stages of a CNN. And the widely-used detector was faced with the dilemma of the dependence of ImageNet pre-trained models. To handle this problem we aggregate a complementary path from the input to pyramid features; the resulting model, ZoomNet, achieves a notable improvement on small objects. Extensive experimental results on two challenging datasets demonstrate the effectiveness of the proposed model. Moreover, our work reveals the gap between classification and detection tasks, and sheds light on handling limitations of regular CNN detectors, which are biased to some specific scales and can not handle extreme scale invariant. And further work may include exploring how to build the extra complementary path so that it can be more efficient, and bridge the gap within classification and detection tasks.

## Acknowledgments

This work was supported in part by the Natural Science Foundation of China (Project Number 61521002).

## References

R. Benenson, M. Mathias, Radu T., and L. Van G. Pedestrian detection at 100 frames per second. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2012.

A. Bhattacharyya, M. Fritz, and B. Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2018.

Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *Proceeding of European conference on computer vision (ECCV)*, 2016.

J. Cao, Y. Pang, and X. Li. Exploring multi-branch and high-level semantic networks for improving pedestrian detection. *arXiv preprint arXiv:1804.00872*, 2018.

M. Cordts, M. Omran, S. Ramos, M. Rehfeld, T.and Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2016.

J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Proceeding of Advances in Neural Information Processing Systems (NIPS)*, 2016.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2005.

P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(4):743–761, 2012.

P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(8):1532–1545, 2014.

G. Duan, H. Ai, J. Xing, S. Cao, and S. Lao. Scene aware detection and block assignment tracking in crowded scenes. *Image and Vision Computing (IVC)*, 30(4-5):292–305, 2012.

P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2008.

C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2014.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2016.

K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceeding of IEEE International Conference on Computer Vision (ICCV)*, 2017.

T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: Towards accurate region proposal generation and joint object detection. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2016.

T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen. Ron: Reverse connection with objectness prior networks for object detection. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2017.

J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan. Scale-aware fast r-cnn for pedestrian detection. *arXiv preprint arXiv:1510.08160*, 2015.

X. Li, Z. Jie, W. Wang, C. Liu, J. Yang, X. Shen, Z. Lin, Q. Chen, S. Yan, and J. Feng. Foveanet: Perspective-aware urban scene parsing. *arXiv preprint arXiv:1708.02421*, 2017.

Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun. Detnet: A backbone network for object detection. In *Proceeding of European conference on computer vision (ECCV)*, 2018.

T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. *arXiv preprint arXiv:1612.03144*, 2016.

S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. *arXiv preprint arXiv:1803.01534*, 2018.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Proceeding of European Conference on Computer Vision (ECCV)*, 2016.

J. Mao, T. Xiao, Y. Jiang, and Z. Cao. What can help pedestrian detection? In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2017.

J. Marin, D. Vázquez, A. M López, J. Amores, and B. Leibe. Random forests of local experts for pedestrian detection. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2013.

W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang. Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 40(8):1874–1887, 2018.

J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2016.

S. Ren, R. He, K.and Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceeding of Advances in Neural Information Processing Systems (NIPS)*, 2015.

B. Singh and L. S Davis. An analysis of scale invariance in object detection-snip. *arXiv preprint arXiv:1711.08189*, 2017.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2001.

X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2017.

B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision (IJCV)*, 75(2):247–266, 2007.

J. Xing, H. Ai, and S. Lao. Multiple human tracking based on multi-view upper-body detection and discriminative learning. In *Proceeding of International Conference on Pattern Recognition (ICPR)*, 2010.

J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z Li. Robust multi-resolution pedestrian detection in traffic scenes. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2013.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

L. Zhang, L. Lin, X. Liang, and K. He. Is faster R-CNN doing well for pedestrian detection? In *Proceeding of European Conference on Computer Vision (ECCV)*, 2016a.

S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2015.

S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2016b.

S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2017.

S. Zhang, J. Yang, and B. Schiele. Occluded pedestrian detection through guided attention in cnns. In *Proceeding of IEEE Conference of Computer Vision and Pattern Recognition (CVPR)*, 2018a.

X. Zhang, L. Cheng, B. Li, and H. M. Hu. Too far to see? not really! pedestrian detection with scale-aware localization policy. *IEEE Transactions on Image Processing (TIP)*, 27 (8):3703–3715, 2018b.