Cars Can’t Fly up in the Sky: Improving Urban-Scene Segmentation via Height-driven Attention Networks

Sungha Choi¹,²,³ Joanne T. Kim²,³ Jaegul Choo⁴
¹A&B Center, LG Electronics, Seoul, South Korea
²Lawrence Livermore National Laboratory, Livermore, CA, USA
³Korea University, Seoul, South Korea
⁴KAIST, Daejeon, South Korea
{shachoi, tengyee}@korea.ac.kr jchoo@kaist.ac.kr

Abstract
This paper exploits the intrinsic features of urban-scene images and proposes a general add-on module, called height-driven attention networks (HANet), for improving semantic segmentation for urban-scene images. It emphasizes informative features or classes selectively according to the vertical position of a pixel. The pixel-wise class distributions are significantly different from each other among horizontally segmented sections in the urban-scene images. Likewise, urban-scene images have their own distinct characteristics, but most semantic segmentation networks do not reflect such unique attributes in the architecture. The proposed network architecture incorporates the capability exploiting the attributes to handle the urban-scene dataset effectively. We validate the consistent performance (mIoU) increase of various semantic segmentation models on two datasets when HANet is adopted. This extensive quantitative analysis demonstrates that adding our module to existing models is easy and cost-effective. Our method achieves a new state-of-the-art performance on the Cityscapes benchmark with a large margin among ResNet-101 based segmentation models. Also, we show that the proposed model is coherent with the facts observed in the urban scene by visualizing and interpreting the attention map. Our code and trained models are publicly available.

1. Introduction
Semantic image segmentation, a fundamental task in computer vision, is employed for basic urban-scene understanding in autonomous driving. Fully convolutional networks (FCNs) [28] are seminal work that adopts deep convolutional neural networks (CNNs) in semantic segmentation, by replacing fully connected layers with convolutional ones at the last stage in typical CNN architectures. Other advanced techniques, such as skip-connections in an encoder-decoder architecture [2, 34, 8], an atrous convolution [6, 46], an atrous spatial pyramid pooling (ASPP) [7], and a pyramid pooling module [51], have further improved the FCN-based architecture in terms of semantic segmentation performance. They have proven to be successful in diverse semantic segmentation benchmarks [14, 15, 31, 12, 1, 25, 32] including urban-scene datasets.

Yet, urban-scene images have their own distinct nature related to perspective geometry [23] and positional patterns [55, 10]. Due to the fact that the urban-scene images are captured by the cameras mounted on the front side of a car, the urban-scene datasets consist only of road-driving pictures. This leads to the possibility of incorporating common structural priors depending on a spatial position, markedly in a vertical position. To verify this characteristic, we present the class distribution of an urban-scene dataset across vertical positions in Fig. 1. Although the pixels of the few classes are dominant in an entire region of an image (Fig. 1(a)), the class distribution has significant dependency on a vertical position. That is, a lower part of an image is mainly composed of road, while the middle part contains various kinds of relatively small objects. In the upper part, buildings, vegetation, and sky are principal objects as shown in Fig. 1(b).

Table 1 shows the probabilities with respect to the dominant top-5 classes: road, building, vegetation, car, and sidewalk. The class distribution is extremely imbalanced that the dominant classes take over 88% of the entire dataset. As mentioned above, the class distribution is completely different if the image is divided into three regions: upper, middle, and lower parts. For instance, the probability of the road class \(p_{\text{road}}\) is 36.9% on average given an entire image, but this chance drops dramatically to 0.006% for an upper region, while jumps to 87.9% if a lower region is considered.

Also, we analyzed this observation using entropy, a measure of uncertainty. The entropy of the probability distribution \(X\) of a pixel over 19 classes in the Cityscapes dataset [12] is computed as

\[ H(X) = -\sum_{i=1}^{19} p_i \log p_i \]

\[ H(X) \text{ is high if the class distribution is imbalanced.} \]
Motivation of our approach, the pixel-wise class distributions. All numbers are average values obtained from the entire training set of the Cityscapes dataset with its pixel-level class labels [12]. Note that there exists a total of 2048K pixels per image, and the y-axis is in log-scale. (a) Each bar represents the average number of pixels assigned to each class contained in a single image. For example, on average, about 685K pixels per image are assigned to the road class. (b) Each part of an image divided into three horizontal sections has a significantly different class distribution from each other. For example, the upper region has just 38 pixels of the road class, while the lower region has 480K pixels of it.

\[ H(X) = H(p_{\text{road}}, p_{\text{building}}, \ldots, p_{\text{motorcycle}}) = - \sum_i p_i \log p_i, \]  

where \( p_i \) denotes the probability that an arbitrary pixel is assigned to the \( i \)-th class. The conditional entropy \( H(X|\text{image}) \) of \( X \) given an image, computed by Eq. (1), is 1.84. On the other hand, the average conditional entropy of \( X \) given each of the three regions as \( H(X|\text{upper}) \), \( H(X|\text{middle}) \), and \( H(X|\text{lower}) \), is 1.26 as shown in Table 1. As a result, one can see the uncertainty is reduced if we divide an image into several parts horizontally. Based on this analysis, if we can identify the part of an image to which a given arbitrary pixel belongs, it will be helpful for pixel-level classification in semantic segmentation.

Motivated by these observations, we propose a novel height-driven attention networks (HANet) as a general add-on module to semantic segmentation for urban-scene images. Given an input feature map, HANet extracts “height-wise contextual information”, which represents the context of each horizontally divided part, and then predicts which features or classes are more important than the others within each horizontal part from the height-wise contextual information. The models adopting HANet consistently outperform baseline models. Importantly, our proposed module can be added to any CNN-based backbone networks with negligible cost increase. To verify the effectiveness of HANet, we conduct extensive experiments with various backbone networks such as ShuffleNetV2 [29], MobileNetV2 [37], ResNet-50 [18], and ResNet-101 [18]. We also focus on lightweight backbone networks where a lightweight HANet is more effective. The main contributions of this paper include:

- We propose a novel lightweight add-on module, called HANet, which can be easily added to existing models and improves the performance by scaling the activation of channels according to the vertical position of a pixel. We show the effectiveness and wide applicability of our method through extensive experiments by applying on various backbone networks and two different datasets.
- By adding HANet to the baseline, with negligible computational and memory overhead, we achieve a new state-of-the-art performance on the Cityscapes benchmark with a large margin among ResNet-101 based models.
- We visualize and interpret the attention weights on individual channels to experimentally confirm our intuition and rationale that height position is crucial to improve the segmentation performance on urban scene.

### 2. Related Work

**Model architectures for semantic segmentation.** Maintaining the resolution of a feature map while capturing high-level semantic features is essential in achieving high performance of semantic segmentation. Typically, high-level features are extracted by stacking multiple convolutions and spatial pooling layers, but the resolution gets coarser in the process. Several studies [28, 33] address this limitation by leveraging deconvolution for learnable upsampling.
from low-resolution features. Skip-connections overcome the limitation by recovering the object boundaries in a decoder layer through leveraging high-resolution features existing earlier in the encoder layer [2, 34, 24, 8]. Another prevalent method is atrous convolution [6, 46], which increases the receptive field size without increasing the number of parameters, and it is widely adopted in recent semantic segmentation networks [7, 8, 43, 49, 53]. Additionally, ASPP [7] and pyramid pooling modules [51] address such challenges caused by diverse scales of objects.

More recently, long-range dependency is captured [54, 21, 16] to improve the performance especially by extending self-attention mechanism [39]. Focusing on boundary information is another approach in semantic segmentation [3, 5, 13, 38, 30]. Recent work imposes separate modules for targeting boundary processing [13, 58] or boundary driven adaptive downsampling [30]. Also, capturing contextual information is widely exploited. ACFNet [48] uses class center to gather features of pixels in each class as categorical context, while CFNet [50] learns the distribution of co-occurrent features and captures co-occurrent context to relook before making predictions. CiSS-Net [52] adopts reinforcement learning to explore the context information in predicted segmentation maps, not having any supervision.

**Exploitation of urban-scene image.** In the field of semantic segmentation, several studies exploit the characteristics of the urban-scene images. In general, the scale of objects significantly vary in the urban-scene images. FoveaNet [25] localizes a “fovea region”, where the small-scale objects are crowded, and performs scale normalization to address heterogeneous object scales. DenseASPP [43] adopts densely connected ASPP, which connects multiple atrous convolutional layers [20] to address large-scale changes of the objects. Another recent approach [53] exploits the fact that the urban-scene images have continuous video frame sequences and proposes the data augmentation technique based on a video prediction model to create future frames and their labels.

Recent approaches in the field of domain adaptation propose the method to leverage the properties of the urban-scene images. A class-balanced self-training with spatial priors [55] generates pseudo-labels for unlabeled target data by exploiting which classes appear frequently at a particular position in an image for unsupervised domain adaptation. Another approach [10] divides an urban-scene image into several spatial regions and conducts domain adaptation on the pixel-level features from the same spatial region. Also, the correlation between depth information and semantic is exploited to gain additional information from synthetic data for urban-scene domain adaptation [9].

**Channel-wise attention.** Our proposed method, HANet, has strong connections to a channel-wise attention approach, which exploits the inter-channel relationship of features and scales the feature map according to the importance of each channel. Squeeze-and-excitation networks (SENets) [19] capture the global context of the entire image using global average pooling and predict per-channel scaling factors to extract informative features for an image classification task. This mechanism is widely adopted in subsequent studies [40, 49, 44, 22] for image classification and semantic segmentation tasks. Inspired from ParseNet [27], which shows the impact of a global context of an entire image in semantic segmentation, previous work [49, 44, 22] for semantic segmentation exploits the global context of the entire image to generate channel-wise attention. However, the urban-scene datasets [12, 1, 32, 4] consist only of road-driving pictures, which means that the images share similar class statistics. Therefore, the global context should be relatively similar among urban-scene images. As a result, the global context of the entire image cannot present distinct information of each image to help per-pixel classification in urban-scene images. This explains why the previous work related to channel-wise attention for semantic segmentation mainly focuses on the generic scene datasets.

### 3. Proposed Method

Urban-scene images generally involve common structural priors depending on a spatial position. Each row of an image has significantly different statistics in terms of a category distribution. In this sense, individually capturing the height-wise contextual information, which represents the global context of each row can be used to estimate how channels should be weighted during pixel-level classification for urban-scene segmentation. Therefore, we propose HANet which aims to i) extract the height-wise contextual information and ii) compute height-driven attention weights to represent the importance of features (at intermediate layers) or classes (at last layer) for each row using the context.

In this section, we first describe HANet as a general add-on module and then present the semantic segmentation networks incorporating several HANet at different layers specialized for urban-scene segmentation.

#### 3.1. Height-driven Attention Networks (HANet)

HANet generates per-channel scaling factors for each individual row from its height-wise contextual information as its architecture illustrated in Fig. 2.

Let $X_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ and $X_h \in \mathbb{R}^{C_h \times H_h \times W_h}$ denote the lower- and higher-level feature maps in semantic segmentation networks, where $C$ is the number of channels, $H$ and $W$ are the spatial dimensions of the input tensor, height and width, respectively. Given the lower-level feature map $X_i$, $F_{HANet}$ generates a channel-wise attention map $A \in \mathbb{R}^{C_h \times H_h}$ made up of height-wise per-channel scaling factors and fitted to the channel and height dimensions of the higher-level feature map $X_h$. This is done in a series of steps: width-wise pooling (Fig. 2(a)), interpolation for
coarse attention (Fig. 2(b,d)), and computing height-driven attention map (Fig. 2(c)). Moreover, adding positional encoding is included in the process (Fig. 2(e)).

After computing the attention map, the given higher-level feature map $X_h$ can be transformed into a new representation $\hat{X}_h$, acquired by an element-wise multiplication of $A$ and $X_h$. Note that single per-channel scaling vector is derived for each individual row or for each set of multiple consecutive rows, so the vector is copied along with the horizontal direction, which is formulated as

$$\hat{X}_h = F_{\text{HANet}}(X_e) \odot X_h = A \odot X_h.$$  

(2)

**Width-wise pooling (Fig. 2(a)).** In order to obtain a channel-wise attention map, we firstly extract height-wise contextual information from each row by aggregating the $C_L \times H_L \times W_L$ input representation $X_e$ into a $C_L \times H_L \times 1$ matrix $Z$ using a width-wise pooling operation $G_{pool}$, i.e.,

$$Z = G_{pool}(X_e).$$  

(3)

There are two typical pooling methods, average pooling and max pooling, to squeeze the spatial dimension. The choice between max pooling and average pooling for the width-wise pooling operation is a hyper-parameter and is empirically set to average pooling. Formally, the $h$-th row vector of $Z$ is computed as

$$Z_{:,h} = \left[ \frac{1}{W} \sum_{i=1}^{W} X_{1,h,i}; \ldots; \frac{1}{W} \sum_{i=1}^{W} X_{C,h,i} \right].$$  

(4)

**Interpolation for coarse attention (Fig. 2(b,d)).** After the pooling operation, the model generates a matrix $Z \in \mathbb{R}^{C_L \times H_L}$. However, not all the rows of matrix $Z$ may be necessary for computing an effective attention map. As illustrated in Fig. 1(b), class distributions for each part highly differ from each other, even if we divide the entire area into just three parts. Therefore, we interpolate $C_L \times H_L$ matrix $Z$ into $C_L \times H$ matrix $\hat{Z}$ by downsampling it (Fig. 2(b)). $\hat{H}$ is a hyper-parameter and is empirically set to 16. Since the attention map, constructed from downsampled representations, is also coarse, the attention map is converted to have the equivalent height dimension with the given higher-level feature map $X_h$ via upsampling (Fig. 2(d)).

**Computation of height-driven attention map (Fig. 2(c)).** A height-driven channel-wise attention map $A$ is obtained by convolutional layers that take width-wise pooled and interpolated feature map $\hat{Z}$ as input. Recent work that utilized a channel-wise attention in classification and semantic segmentation [19, 40, 49] adopts fully connected layers rather than convolutional layers since they generate a channel-wise attention for an entire image. However, we adopt convolutional layers to let the relationship between adjacent rows be considered while estimating the attention map since each row is related to its adjacent rows.

The attention map $A$ indicates which channels are critical at each individual row. There may exist multiple informative features at each row in the intermediate layer; in the last layer, each row can be associated with multiple labels (e.g., road, car, sidewalk, etc.). To allow these multiple features and labels, a sigmoid function is used in computing the attention map, not a softmax function. These operations consisting of $N$ convolutional layers can be written as

$$A = G_{\text{up}} \left( \sigma \left( G_{\text{conv}}^N \left( \ldots \left( \sigma \left( G_{\text{conv}}^1 \left( \delta \left( G_{\text{conv}}^1 \left( Z \right) \right) \right) \right) \right) \right) \right),$$  

(5)

where $\sigma$ is a sigmoid function, $\delta$ is a ReLU activation, and $G_{\text{conv}}$ denotes $i$-th one-dimensional convolutional layer. We empirically adopt three convolutional layers: the first one $G_{\text{conv}}^1 = Q^1 \in \mathbb{R}^{C_L \times \hat{H}}$ for channel reduction, the second one $G_{\text{conv}}^2 (\delta(Q^1)) = Q^2 \in \mathbb{R}^{2 \cdot C_L \times \hat{H}}$, and the last one $G_{\text{conv}}^3 (\delta(Q^2)) = \hat{A} \in \mathbb{R}^{C_L \times \hat{H}}$ for generating an attention map. The reduction ratio $r$ reduces the parameter overhead of HANet as well as gives a potential regularization effect. An analysis on the effect of various reduction ratio as a hyper-parameter will be presented in Section 4.2.2.

**Incorporating positional encoding (Fig. 2(e)).** When humans recognize a driving scene, they have prior knowledge on the vertical position of particular objects (e.g., road and sky appear in the lower and upper part, respectively). Inspired by this observation, we add the sinusoidal positional
encodings [39] to the intermediate feature map $Q^i$ at the $i$-th layer in the HANet. A hyper-parameter $i$ is analyzed in the supplementary material. For injecting positional encodings, we follow the strategy proposed in Transformer [39]. The dimension of the positional encodings is same as the channel dimension $C$ of the intermediate feature map $Q^i$. The positional encodings are defined as

$$PE_{(p, 2i)} = \sin \left( \frac{p}{100^{2i/C}} \right)$$

$$PE_{(p, 2i+1)} = \cos \left( \frac{p}{100^{2i/C}} \right),$$

where $p$ denotes the vertical position index in the entire image ranging from zero to $H - 1$ of coarse attention, and $i$ is the dimension. The number of the vertical position is set to $H$ as the number of rows in coarse attention. The new representation $\hat{Q}$ incorporating positional encodings is formulated as

$$\hat{Q} = Q \oplus PE,$$

(6)

where $\oplus$ is an element-wise sum.

Height positions are randomly jittered by up to two positions to generalize over different camera location from various datasets to prevent an inordinately tight position-object coupling. Additionally, we experiment with using learnable positional embeddings [17] to find the best way to incorporate positional information in the supplementary material.

Meanwhile, CoordConv [26] proposed to embed height and width coordinates in the intermediate features for various vision tasks: extra channels containing hard-coded coordinates (e.g., height, width, and optional $r$) are concatenated channel-wise to the input representation, and then a standard convolutional layer is applied. Unlike this model, HANet exploits positional information of height to obtain attention values, $C \times H \times 1$, which is used for gating the output representations of main networks. Therefore, HANet differs significantly from CoordConv in terms of how to exploit the positional information. We experimentally compare ours with CoordConv in Section 4.2.2.

### 3.2. Segmentation Networks based on HANet

We adopt DeepLabv3+ [6] as a baseline for semantic segmentation. DeepLabv3+ has an encoder-decoder architecture with ASPP that employs various dilation rates. We add HANet to the segmentation networks at five different layers (Fig. 3) after the point where high-level representation is encoded from backbone networks. This is because the higher-level feature has a stronger correlation with the vertical position. We conduct an ablation study to see the performance gain from adding HANet at different layers.

### 3.3. Comparison with Other Attention Strategies

Self-attention-based approaches like DANet [16], obtain attention values, $(H \times W) \times (H \times W), C \times C$, from the semantic interdependencies in spatial and channel dimensions, respectively. However, HANet does not consider the interdependencies among the dimensions. HANet includes one-dimensional convolutional layers being a separate branch as a modular design (Fig. 2(a)-(d)), which specifically consider the structural property in urban-scene data. In this manner, HANet derives attention values, $C \times H \times 1$, to gate activation at the horizontal section of a feature map output in the main networks, considering vertical position. HANet is significantly more lightweight than self-attention-based approaches that considers the relationship between every pair in each dimension. Meanwhile, channel-wise attention approaches such as SENet [19] generates attention values, $C \times 1 \times 1$, only at an image level. This is not ideal for urban-scene segmentation as most of the images shares similar circumstances and context.

### 4. Experiments

In this section, we first describe the implementation details of HANet. Then, we experimentally demonstrate the effectiveness and wide applicability of our proposed methods by extensive quantitative analysis including ablation studies. We evaluate HANet on two different urban-scene datasets including Cityscapes [12] and BDD100K [45]. Furthermore, we visualize and analyze the attention map generated from HANet. For all the quantitative experiments, we measure the segmentation performance in terms of mean Intersection over Union (mIoU) metric.

#### 4.1. Experimental Setup

**Base segmentation architecture** Our network architecture for semantic segmentation is based on DeepLabv3+ [8]. We adopt various backbone networks including ShuffleNetV2 [29], MobileNetV2 [37], and ResNet [18] to verify wide applicability of HANet. Note that HANet can be easily inserted on top of various backbone networks. The adopted backbone networks are pretrained on ImageNet [36] for all the experiments, unless otherwise noted.

**Stronger baseline** To strictly verify the effectiveness of HANet, we improved the performance of the DeepLabv3+ baseline adopting ResNet-101, by integrating SyncBatch-
Table 2: Comparison of mIoU, the number of model parameters and FLOPs between the baseline and HANet on Cityscapes validation set according to various backbone networks and output stride. Adding HANet to the baseline consistently increase the mIoU with minimal cost increase.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>OS</th>
<th>Models</th>
<th>Params (M)</th>
<th>GFLOPs</th>
<th>mIoU(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNet V2 [29]</td>
<td>16</td>
<td>Baseline</td>
<td>14.8M</td>
<td>428.70</td>
<td>73.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+HANet</td>
<td>16.1M</td>
<td>428.82</td>
<td>74.70</td>
</tr>
<tr>
<td>ResNet-50 [18]</td>
<td>16</td>
<td>Baseline</td>
<td>45.1M</td>
<td>553.74</td>
<td>76.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+HANet</td>
<td>47.6M</td>
<td>553.85</td>
<td>77.78</td>
</tr>
<tr>
<td>ResNet-101 [18]</td>
<td>16</td>
<td>Baseline</td>
<td>64.2M</td>
<td>765.53</td>
<td>77.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+HANet</td>
<td>65.4M</td>
<td>765.63</td>
<td>79.31</td>
</tr>
</tbody>
</table>

4.2. Cityscapes

The Cityscapes dataset [12] is a large-scale urban-scene dataset, holding high-quality pixel-level annotations of 5K images and 20K coarsely annotated images. Finely annotated images consist of 2,975 train images, 500 validation images, and 1,525 test images. The annotations of test images are withheld for benchmarks. The resolution of each image is 2048×1024, and 19 semantic labels are defined. In all the experiments on Cityscapes validation set, we train our models using finely annotated training set for 40K iterations with a total batch size of 8 and a crop size of 768×768.
shows the superior results of HANet in ResNext-101 [78.52 ±0.7] compared to other models. When +HANet is used, the performance increases by +1.60% on average. This implies the performance improvement in the entire region. Indeed, the performance significantly rises on the upper and lower regions as in Table 5.

### 4.2.4 Efficacy at segmented regions
As mentioned in Section 1, the average entropy decreases as we divide the image into multiple horizontal subsections. This implies the performance improvement in the entire region of an image. Besides, the entropy of the upper and lower regions have low entropy compared to the middle region. In this respect, we expect the performance increase arising from HANet would be larger in the upper and lower regions than that in the middle or entire region. Therefore, we divided images into four horizontal sections and performed ablation studies and hyper-parameter impacts with respect to the HANet injected layers, using positional encodings or not, and channel reduction ratio. ResNet-101, output stride 16 on Cityscapes validation set.

### 4.2.5 Comparison to other state-of-the-art models
To compare with other state-of-the-arts, we train our models using finely annotated training and validation set for 90K iterations. In case of adopting ResNet-101 [42] backbone, coarsely annotated images are additionally used and the model is pretrained on Mapillary [32]. The crop size and batch size are changed into 864×864 and 12, respectively. The inference techniques from Section 4.2.3 are also used, but we do not adopt any other techniques such as online bootstrapping for hard training pixels [41]. We compare our best models based on ResNet-101 and ResNext-101 with other recent models on the Cityscapes test set (Table 6). Our models achieve a new state-of-the-art performance.

### 4.3. BDD100K
Berkeley Deep Drive dataset (BDD100K) [45] is a large-scale diverse driving video database. It includes a semantic image segmentation dataset, consisting of 7,000 training and 1,000 validation images with a resolution of 1280×720. It is a challenging dataset including images of various driving circumstances such as day, night, and diverse weather conditions. Table 7 shows the superior results of HANet in BDD100K. The training strategy of BDD100K is similar to Cityscapes, but we change the crop size into 608×608 and train for 60K iterations with a total batch size of 16. These adjustments simply comes from smaller image size and larger dataset size compared to Cityscapes.

### 4.4. Qualitative Analysis
**Attention map visualization.** We visualize the attention weights to analyze the behavior of the proposed HANet. The attention visualization highlights those channels emphasized at a different vertical position. Through the visualization, we can find out interesting clues to validate our model.
Table 7: Comparison of mIoU between the baseline and HANet on BDD100K validation set. The output stride is set to 16 and the crop size is 608×608.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>OS</th>
<th>Baseline</th>
<th>+HANet</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNetV2</td>
<td>16</td>
<td>58.91%</td>
<td>60.05%</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>16</td>
<td>63.75%</td>
<td>64.56%</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>8</td>
<td>64.84%</td>
<td>65.60%</td>
</tr>
</tbody>
</table>

Fig. 5 clearly shows that HANet assigns a different amount of attention to a different vertical position, indicating that the model properly learns structural priors with respect to the height in urban-scene data.

![Height-wise class distributions and attention map visualization](image)

**Figure 5:** Visualization of attention map from HANet at different layers. x-axis denotes the channels, and y-axis denotes the height of the feature map, showing which channels are weighted at different vertical position. The channels focused by each height are clearly different. To better visualize, the channels are clustered.

Fig. 6 visualizes the attention map from HANet L2, which computes the attention weights for ASPP layer. In ASPP layer, the channels are obtained from convolution filters that have multi-scale receptive fields and grouped for each scale. Therefore, we can interpret the HANet L2 attention map by each group, with the sequence of channels remaining unchanged. Colored boxes give us insights of our method and the urban-scene images. The green boxes in Fig. 6 shows the low-focused channels of the middle region of images, while blue box indicates the channels which are relatively more focused. That is, the channels obtained from the small receptive field are weighted in the middle region. Since the middle region is where small objects are crowded as pointed in the left figure in Fig. 6, small receptive fields are effective for this region and vice versa. In this manner, we verify that HANet properly learns and captures the intrinsic features of urban scene.

Fig. 7 illustrates that the distribution of the attention map (right figure) from HANet at the last layer, which is following the actual height-wise class distribution (left figure) obtained from the Cityscapes training images. Each class gives a different weight according to the vertical position, meaning that the model actually uses vertical positions in the image. This information corresponds to the observation we introduced through the class distribution analysis in Fig. 1 and Table 1. For instance, road (class 0) appears only in the middle and the lower regions, while sky (class 10) is mainly emphasized in the upper rows.

Visualization of semantic segmentation results are shown in the supplementary material.

**5. Conclusions**

In this paper, we proposed HANet for urban-scene segmentation as an effective and wide applicable add-on module. Our method exploits the spatial priors existing in urban-scene images to construct a cost-effective architecture. We demonstrated the performance increase by adding our method to the baseline model with negligible cost increase. Moreover, we visualized and analyzed the attention maps to show the validity of our initial hypothesis that exploiting vertical positional information helps for urban-scene semantic segmentation.

**Acknowledgments** This work was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)), the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. NRF-2018M3E3A1057305), and the National Supercomputing Center with supercomputing resources including technical support (KSC-2019-INO-0011).
References


