ACNET: ATTENTION BASED NETWORK TO EXPLOIT COMPLEMENTARY FEATURES FOR RGBD SEMANTIC SEGMENTATION

Xinxin Hu, Kailun Yang, Lei Fei and Kaiwei Wang

College of Optical Science and Engineering, Zhejiang University

ABSTRACT

Compared to RGB semantic segmentation, RGBD semantic segmentation can achieve better performance by taking depth information into consideration. However, it is still problematic for contemporary segmenters to effectively exploit RGBD information since the feature distributions of RGB and depth (D) images vary significantly in different scenes. In this paper, we propose an Attention Complementary Network (ACNet) that selectively gathers features from RGB and depth branches. The main contributions lie in the Attention Complementary Module (ACM) and the architecture with three parallel branches. More precisely, ACM is a channel attention-based module that extracts weighted features from RGB and depth branches. The architecture preserves the inference of the original RGB and depth branches, and enables the fusion branch at the same time. Based on the above structures, ACNet is capable of exploiting more high-quality features from different channels. We evaluate our model on SUN-RGBD and NYUDv2 datasets, and prove that our model outperforms state-of-the-art methods. In particular, a mIoU score of 48.3% on NYUDv2 test set is achieved with ResNet50. We will release our source code based on PyTorch and the trained segmentation model at https://github.com/anheidelonghu/ACNet.

Index Terms— Attention, Complementary, RGBD semantic segmentation

1. INTRODUCTION

Semantic segmentation is a basic task of computer vision, whose purpose is to partition an image into several coherent semantically-meaningful parts. Compared with traditional approaches that need to be deployed in complex separate ways, semantic segmentation can be utilized to unify diverse detection tasks desired by navigation systems, at least in standard outdoor conditions [1][2].

In contrast, indoor semantic segmentation that has not been thoroughly investigated, remains challenging in several
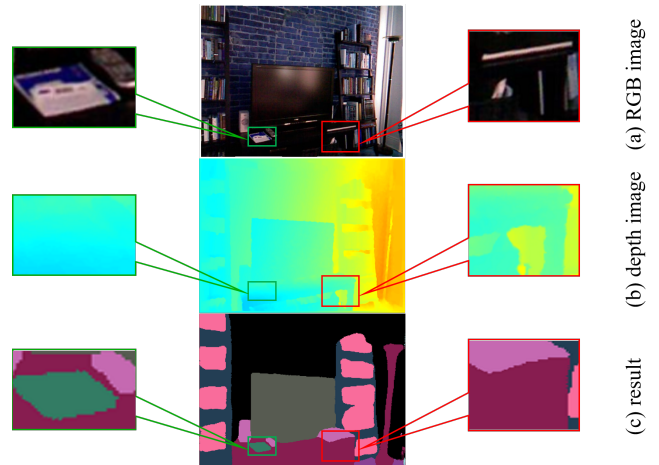
Fig. 1. RGB and depth images have different distributions of features that are appropriately exploited by our ACNet.

aspects. For example, it suffers from uneven illumination and messy spatial overlapping. With the emergence and development of RGBD cameras (such as RealSense, Kinect, Xition etc.), indoor semantic segmentation can be benefited from RGBD observations that encode real-world geometric information, which theoretically leads to better segmentation performance compared to RGB semantic segmentation. Towards this end, there were a few attempts like [3][4] that treated depth image as an additional channel, and used the method similar to RGB semantic segmentation to implement RGBD semantic segmentation. In [5], two neural networks branches were designed for RGB input and depth input, which were merged before upsampling. In [3], depth images was decomposed into three channels, namely disparity, height and angle which were also treated as an RGB image. In [6], instead of simply using traditional convolution neural networks, novel graph-based networks were applied to excavate scene geometric information more sufficiently. In [7], traditional convolution was modified according to depth values. These networks designed for RGBD semantic segmentation have achieved break-through results. However, there are still some issues that need to be solved:

- Although the geometric information encoded in the depth image can clearly provide additional benefits for image segmentation, the information contained in
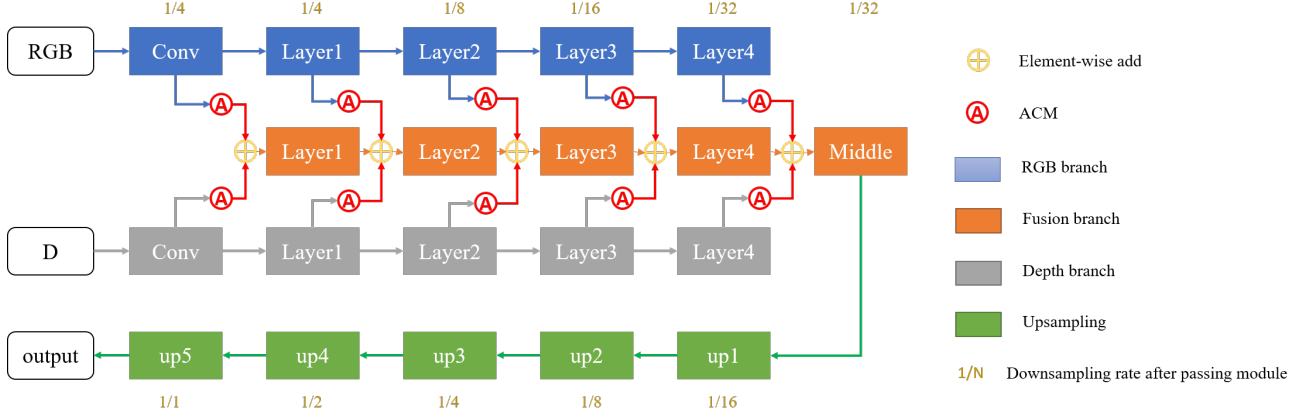
Fig. 2. The overview of our proposed ACNet. RGB image and depth image are processed by two ResNets separately. Red arrows represent the data flow of feature maps reorganized by ACM.

RGB image and depth image are not equivalent for each scene (shown in Fig. 1). In other words, features extracted from RGB branch and depth branch by current networks may be not appropriate.

- Conventional RGBD segmentation network can be divided into two types of architectures. One of them, such as [8], employs two encoders to extract features from RGB and depth image respectively, and combines the features of both before or during upsampling. The other like [5][9] just fuses the RGBD features at the downsampling stage. The former can't sufficiently combine RGBD information, and the latter tends to lose original RGB and depth branches since the fusion branches take the place of them.

In this paper, we propose ACNet (shown in Fig. 2) to combine RGB and depth features by a proportion determined by the input. In ACNet, there are two independent branches based on ResNet [10] to extract features for RGB and depth image separately. Several Attention Complementary Module (ACMs) are designed to obtain features from the aforementioned branches, which are determined by the amount of information they carry. There's another branch based on ResNet to process the merged features. The proposed architecture is able to keep original RGBD features flow as well as to utilize merged features in an integrated network.

2. FRAMEWORK

Attention Complementary Module (ACM). As illustrated in Fig. 1, the information contained in RGB image and depth image vary in different regions of indoor scenes. In order to gather features selectively from RGB branch and depth branch, we have designed a set of attention modules [11] to make the network focus on more informative regions. More precisely, the proposed ACM is based on channel attention [12] (shown in Fig. 3). Assuming the input feature maps

$A = [A_1, \cdots, A_C] \in \mathbb{R}^{C \times H \times W}$, we first apply global average pooling, to have the output $Z \in \mathbb{R}^{C \times 1 \times 1}$, where C denotes the number of channels, H, W denote the height and width of feature maps respectively. The k-th ($k \in [1, C]$) of Z can be expressed as:

$$Z_k = \frac{1}{H \times W} \sum_i^H \sum_j^W A_k(i, j) \tag{1}$$

Then Z is reorganized by a 1×1 convolution layer with the same number of channels as Z. A 1×1 convolution layer is able to excavate correlations between channels, thus eliciting an appropriate weight distribution for these channels. A sigmoid function is applied to activate the convolution result, constraining the value of weight vector $V \in \mathbb{R}^{C \times 1 \times 1}$ between 0 and 1. Finally, we perform an outer product for A and V, and the result $U \in \mathbb{R}^{C \times H \times W}$ can be expressed as:

$$U = A \otimes \sigma[\phi(Z)] \tag{2}$$

where $\otimes$ denotes outer product, σ denotes sigmoid function, and ϕ denotes 1×1 convolution. In this way, feature maps U are converted into new feature maps U, which contain more valid information.

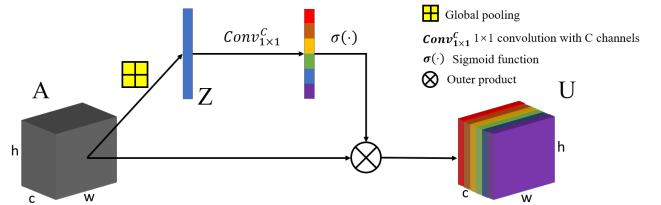**Fig. 3**. Attention Complementary Module (ACM).

Architecture for Feature Fusion. A majority of state-of-the-art RGBD semantic segmentation networks use an encoder that fuses RGBD features either too early or too late [5][8], which ruins the original RGB and depth information or results in a low efficiency of exploiting the carried information. In order to keep the original RGB and depth
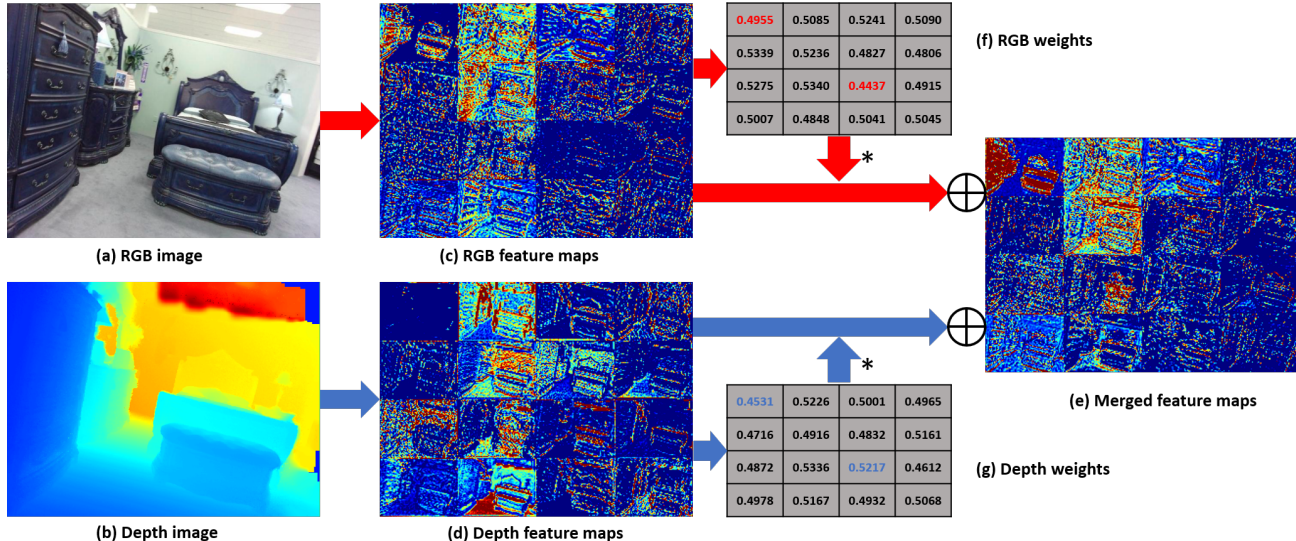
0.4955	0.5085	0.5241	0.5090
0.5339	0.5236	0.4827	0.4806
0.5275	0.5340	0.4437	0.4915
0.5007	0.4848	0.5041	0.5045

(f) RGB weights

0.4531	0.5226	0.5001	0.4965
0.4716	0.4916	0.4832	0.5161
0.4872	0.5336	0.5217	0.4612
0.4978	0.5167	0.4932	0.5068

(g) Depth weights

(a) RGB image (c) RGB feature maps (e) Merged feature maps (b) Depth image (d) Depth feature maps

Fig. 4. How ACM fuses complementary RGBD features into fusion branch. $*$ denotes element-wise product and $\oplus$ denotes element-wise add. The feature maps are visualized from layer2. (f) and (g) depict the weights calculated from the feature maps by ACM, which are multiplied to feature maps separately, and added into the merged features from the fusion branch.

features flow during downsampling, we propose a specialized architecture for RGBD feature fusion. As illustrated in Fig. 2, two complete ResNets are deployed to extract RGB and depth features separately. Note that here the ResNet can be replaced with other networks, e.g., ERF-PSPNet [2] in efficiency-critical domains. Vitally, these two branches can preserve RGB and depth features before upsampling. After that, the fusion branch is leveraged to extract features from the merged feature maps.

Attention Complementary Network (ACNet). We design an integrated network called ACNet for RGBD semantic segmentation. The backbone of ACNet is shown in Fig. 2. RGB image and depth image are inputted, and are processed by ResNet branches separately. During inference, each aforementioned branch provides a group of feature maps at every module stage, such as Conv, Layer1, etc. Then the feature maps are reorganized by ACM. After passing through Conv, the feature maps are further element-wisely added as input of fusion branch, while others are added to the output of fusion branch. In this way, both low-level and high-level features can be extracted, reorganized and fused by our ACNet. As for upsampling, we apply the skip connection like [5], which appends the features in downsampling to upsampling with a quite low computation cost.

3. EXPERIMENTS

We evaluate our method on two public datasets:

NYUDv2 [13]: The NYU-Depth V2 data set (NYUDv2) contains 1,449 RGBD images with dense pixel-wise annotation. We divide the dataset into 795 training images and 654 testing images according to the official setting. We use the version with annotations on 40 classes (common ones used in

the literature).

SUN-RGBD [14]: We use SUN-RGBD V1 which have 37 categories and contains 10,335 RGBD images with dense pixel-wise annotations, 5,285 images for training and 5,050 for testing.

As for metrics, we use the prevailing mean Intersection-over-Union over all classes (mIoU) to evaluate the performance of different semantic segmenters.

Implementation Details. As for data augmentation, we apply random scaling, cropping and flipping to both RGB and depth images, and normalize them separately. For RGB images, we also randomly change their color in HSV space. For all the experiments, we use ResNet50 as the encoder, which is pre-trained on ImageNet [15]. Since depth image contains one channel, we average the three channels in ResNet50's first layer to one channels for the depth branch. We use focal loss [16] with the focusing parameter $\gamma = 2$ to supervise the training of our network. During training stage, we also calculate the average loss of outputs from up1 to up5 in Fig. 2 to better optimize our network. During testing stage, we only evaluate the metrics of last output to ensure consistency with the state of the art. We use SGD optimizer with initial learning rate 0.002, momentum 0.9 and weight decay 0.004. Batch size is set to 4 when training on one NVIDIA TITAN Xp. The learning rate is multiplied by 0.8 for every 20 iterations on SUN-RGBD and 100 iterations for NYUDv2.

Analysis of the ACM. To understand ACM better, we visualize the feature maps from layer2 (shown in Fig. 4) since layer2's low-level features are more consistent with visual intuitions. Note that we only visualize the first 16 of 128 feature maps for better illustration. Regarding the weights as matrices starting from (0,0), they correspond to the features maps one by one, where we can find some insightful results. At
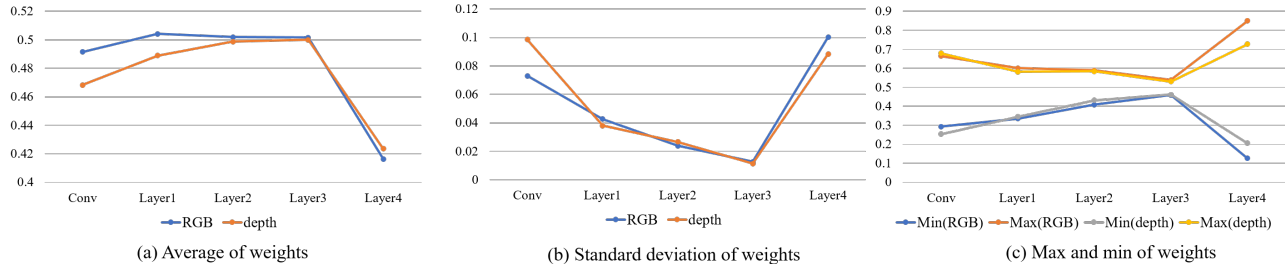
(a) Average of weights	(b) Standard deviation of weights	(c) Max and min of weights

Fig. 5. Quantitative analysis of all ACM in our model

(0,0), feature map of RGB branch contains more valid information than the feature map from depth branch visually, so that ACM tends to give a higher weight to the RGB branch. In contrast, at (2,2), feature map of depth branch contains more information, therefore, depth branch gets higher weight. Finally, feature maps of the two branches are element-wisely added into feature maps of fusion branch to supplement the RGBD information.

We also evaluate weights generated by ACM at all levels (shown in Fig. 5). First, we focus on averages of weights, which indicate the amount of information contained in the feature maps before inputting to ACM. The averages of RGB branches' weights are higher than those of depth branches in Conv and Layer1. This reveals that RGB branches always contain more valid information at lower levels since RGB images often contain many redundant textures. In addition, weights of the two branches in Layer2, Layer3 and Layer4 are queie close, which means RGB and D branches contain nearly equivalent valid information at higher levels. In addition, the average weights of Layer4 are quite low, indicating that the fusion branch might gather enough features. The metrics: std (standard deviation), min and max can reflect the distribution of information across channels. From Conv to Layer3, the std becomes increasingly smaller, which indicates that ACNet can homogenize the distribution of information. However, Layer4 should decide to select useful features as well as eliminate redundant features as it is the last module in the encoder. Therefore, the std of Layer4 is quite high. This experiment demonstrates that our network flattens the distribution of information across channels where complementary features are effectively exploited, which is essential for RGBD semantic segmentation.

Ablation Study. To verify functionality of both ACM and the multi-branch architecture, we perform an ablation study by comparing the original model with two defective models: **Model-1** and **Model-2**. In **Model-1**, we remove all ACMs and the RGB and D branches after Conv Layer. In **Model-2**, we remove all ACMs but retain the multi-branch architecture. Our ablation study on NYUDv2 turns out that, the mIoU of **Model-1** and **Model-2** are 44.3% and 46.8%, verifying the multi-branch architecture and ACM lead to significant accuracy boost of 2.5% and 1.5%, respectively.

Comparison with state-of-the-art networks. We com-

Table 1. Comparison with other state-of-the-art methods on NYUDv2 test set and SUN-RGBD test set.

Model	NYUDv2	SUN-RGBD
3DGNN [6]	39.9%	44.1%
RefineNet (ResNet152) [17]	46.5%	45.9%
Depth-aware CNN [7]	43.9%	42.0%
LSD [8]	45.9%	-
CFN (VGG-16) [18]	41.7%	42.5%
CFN (RefineNet-152) [18]	47.7%	**48.1%**
ACNet (ResNet-50)	**48.3%**	**48.1%**

pare our ACNet with state-of-the-art methods to prove its effectiveness. Note we adopt the most universally-used mIoU as the evaluation metric.

Table 1 shows the result of our ACNet on NYUDv2 and SUN-RGBD test sets. The result shows that on NYUDv2, our ACNet outperforms other state-of-the-art models by 0.6%, yielding the new record of mIoU accuracy 48.3% on NYUDv2. On SUN-RGBD, our model (ResNet-50) is able to reach the same mIoU as CFN (RefineNet-152) [18] by using a more lightweight backbone.

4. CONCLUSIONS

In this paper, we propose a novel multi-branch attention based network for RGBD semantic segmentation. The multi-branch architecture is able to gather features efficiently and doesn't destroy original RGB and depth branches' inference. The attention module can selectively gather features from RGB and depth branches according to the amount of information they contain, and complement the fusion branch by using these weighted features. Our model can resolve the problem that RGB images and depth images always contain unequal amount of information as well as different context distributions. We evaluate our model on NYUDv2 and SUN-RGBD datasets, and the experiments show that our model can outperform state-of-the-art methods.

In the future, we will explore ways to improve the real-time performance of pixel-wise image segmentation not only for RGBD semantic cognition but also for panoramic annular surrounding perception.

5. REFERENCES

[1] Kailun Yang, Luis M Bergasa, Eduardo Romera, Ruiqi Cheng, Tianxue Chen, and Kaiwei Wang, "Unifying terrain awareness through real-time semantic segmentation," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1033–1038.

[2] Kailun Yang, Xinxin Hu, Luis M Bergasa, Eduardo Romera, Xiao Huang, Dongming Sun, and Kaiwei Wang, "Can we pass beyond the field of view? panoramic annular semantic segmentation for real-world surrounding perception," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1–8.

[3] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision*. Springer, 2014, pp. 345–360.

[4] Yang He, Wei-Chen Chiu, Margret Keuper, Mario Fritz, and Saarland Informatics Campus, "Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling.," in *CVPR*, 2017, pp. 7158–7167.

[5] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang, "Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation," *arXiv preprint arXiv:1806.01054*, 2018.

[6] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5199–5208.

[7] Weiyue Wang and Ulrich Neumann, "Depth-aware cnn for rgb-d segmentation," *arXiv preprint arXiv:1803.06791*, 2018.

[8] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang, "Localitysensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 3.

[9] Hao Chen and Youfu Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3051–3060.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, vol. 7, 2017.

[12] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision, Munich, Germany*, 2018, pp. 8–14.

[13] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.

[14] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[16] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[17] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5168–5177.

[18] Di Lin, Guangyong Chen, Daniel Cohen-Or, Pheng-Ann Heng, and Hui Huang, "Cascaded feature network for semantic segmentation of rgb-d images," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1320–1328.