Comparative Analysis of Semantic Segmentation Models for Fisheye Automotive Images

Andrea Marinelli
andrea.marinelli@studenti.unipd.it

Gianmarco Betti
gianmarco.betti@studenti.unipd.it

Abstract

Semantic segmentation is a critical computer vision technique, particularly significant in autonomous driving applications where accurate environment perception is essential. This study investigates the performance of two pre-trained semantic segmentation models, DenseASPP and Gated-SCNN, on the WoodScapes dataset, which contains fisheye images characterized by significant radial distortion. We employ transfer learning to adapt these models, originally trained on the Cityscapes dataset, to effectively manage the distortions present in WoodScapes. The study provides insights into the effectiveness of different architectural components in addressing the challenges posed by distorted images in automotive contexts.

1. Introduction

Semantic segmentation is a computer vision technique that assigns a class label to each pixel in an image, aiming to partition the image into mutually exclusive subsets, in which each subset represents a meaningful region of the original image. This detailed classification turns visual inputs into categorized segments, essential for applications such as autonomous driving. As self-driving cars need to understand their surroundings accurately to navigate safely, the precision of semantic segmentation models is vital.

This study focuses on comparing the performance of two advanced semantic segmentation models, namely DenseASPP and Gated-SCNN, when applied to fisheye automotive images. Fisheye cameras, widely used in automotive systems, introduce significant radial distortions that challenge conventional segmentation models. By leveraging the WoodScapes dataset, known for its fisheye images, this research employs transfer learning and fine-tuning to adapt models pre-trained on the Cityscapes dataset to handle these distortions effectively. Our experiments reveal that the Gated-SCNN model outperforms DenseASPP in terms of mean Intersection over Union (mIoU) and F1-score, demonstrating better boundary precision and class differentiation.

The main goals and contributions of our work include: review and analysis of modern semantic segmentation techniques; investigation of how well pre-trained models can handle fisheye camera images, assessing their robustness and adaptability; exploration of transfer learning techniques, including which layers to freeze during training and how to tune hyperparameters for this task.

2. Related Work

Semantic segmentation on images with radial distortion, such as those captured by fisheye cameras, has traditionally been approached by first correcting the distortion and then applying pre-trained models. This method is limited due to the inherent loss of field of view (FOV) and resampling distortions. Our approach deviates from this conventional method by directly employing models capable of handling distorted images without prior undistortion.

Previous research, as cited in [9], includes works that adapted convolutional neural networks (CNNs) to handle fisheye distortion. For instance, semantic segmentation networks like ENet have been fine-tuned on fisheye datasets, achieving significant improvements by leveraging pixel-wise labels specific to fisheye images. Additionally, the WoodScape dataset itself has been instrumental in exploring these advanced architectures, providing essential benchmarks and facilitating the development of multi-camera and multi-task models that better address the challenges posed by fisheye distortion.

2.1. Semantic Segmentation Techniques

In this section, we examine the evolution of semantic segmentation techniques up to the most recent ones, providing a clear overview of the methodologies used in this field and understanding the context in which our models are situated. This review is based on several comprehensive surveys that have extensively analyzed advancements in semantic segmentation over the years [1] [2] [4] [5].

2.1.1 Historical Perspective

Before deep learning, semantic segmentation used simple methods like thresholding, clustering, and edge detection, which struggled with complex scenarios. Deep learning brought significant advancements with models like Fully Convolutional Networks (FCNs). FCNs replaced fully connected layers with convolutional layers, enabling pixel-wise classification and maintaining spatial information. Region-based approaches like R-CNN adapted for segmentation further improved precision, particularly in tasks requiring detailed boundary understanding.

2.1.2 State-of-the-Art Methods

Recent advancements in semantic segmentation focus on enhancing model capabilities to better handle complex scenes. These advancements can be categorized into:

- **Context-based Methods**: Models like PSPNet and DeepLab use techniques such as atrous convolution and spatial pyramid pooling to aggregate context at multiple scales, capturing both detailed and global scene information. Both DenseASPP and Gated-SCNN leverage these techniques by incorporating Atrous Spatial Pyramid Pooling (ASPP) modules into their architectures.
- **Feature-enhancement-based Methods**: Models like U-Net and RefineNet enhance feature maps through skip connections and multi-path refinement, recovering detailed spatial information lost during downsampling.
- **Attention-based and Transformer Methods**: Recent models incorporate attention mechanisms and transformers, such as SETR and SegFormer, to improve focus on relevant features and handle complex scenes with diverse objects [7].

Each technique brings unique strengths, addressing specific challenges posed by various applications and image conditions.

3. Dataset

WoodScape dataset is an extensive fisheye automotive dataset, named after Robert Wood who invented the fisheye camera in 1906 [9]. Fisheye cameras are designed to capture a wide field of view, making them useful in applications like surveillance, augmented reality, and automotive systems.

This dataset is chosen due to its unique challenges, including significant radial distortions and a wide range of environmental conditions depicted in the images. These characteristics make it an ideal candidate to test the robustness and adaptability of advanced semantic segmentation models.

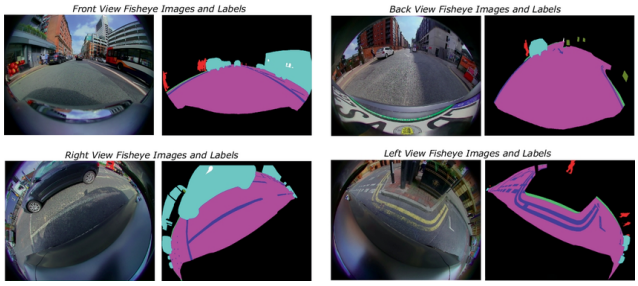

Figure 1. WoodScape dataset image examples and their labels from four views: TOP LEFT: Front-view, TOP RIGHT: Back-view, BOTTOM LEFT: Right-view, BOTTOM RIGHT: Left-view.

3.1. Description

More specifically, this dataset contains high-resolution images (1280 x 960) captured by four surround view cameras (covering 360°). It includes annotations for various tasks such as object detection, semantic segmentation, instance segmentation, and motion segmentation.

3.2. Data preparation

To download the dataset and generate the semantic annotations, we used the scripts provided in the official GitHub repository of the dataset, available at https://github.com/valeoai/WoodScape. Semantic annotations are available for approximately 8200 images. We used the script version that generates annotations for 9 classes plus void, which are: ”road”, ”curb”, ”person”, ”rider”, ”vehicles”, ”bicycle”, ”motorcycle”, and ”traffic sign”. Despite limited resources for training, we decided to use the entire dataset to maximize its benefits and reduce the risk of overfitting.

3.3. Preprocessing

For preprocessing, we applied common data augmentation techniques as proposed in [8]. To avoid overfitting, we used random horizontal flipping, random scaling in the range of [0.5, 2], random brightness jittering within the range of [-10, 10], and random cropping of 483 x 483 image patches. Additionally, to supervise the shape stream in training the GSCNN model, we had to generate ground truth boundaries. This was done using functions provided in the official GSCNN GitHub repository available at https://github.com/nv-tlabs/GSCNN.

4. Models

The choice of DenseASPP and Gated-SCNN for this comparative study was motivated by their innovative use

of advanced architectural elements that address the complexities of semantic segmentation in challenging environments, such as those presented by fisheye images in automotive contexts. Both models incorporate state-of-the-art techniques such as atrous convolution and spatial pyramid pooling, yet they differ significantly in their core architectures and mechanisms, offering a unique comparative perspective. DenseASPP extends the DenseNet architecture with atrous spatial pyramid pooling to enhance its capability in capturing multi-scale context, while Gated-SCNN introduces a gating mechanism to refine feature extraction and segmentation by effectively utilizing shape information.

4.1. DenseASPP

Densely connected Atrous Spatial Pyramid Pooling (DenseASPP), proposed in [8], stands out for its dense connectivity pattern and incorporation of atrous spatial pyramid pooling. This allows the network to cover a wide range of scales and improves its adaptability to various object sizes and shapes encountered in urban scenes. The architecture, shown in Figure 2, effectively exploits the strengths of DenseNet, introduced in [3], known for its efficiency in feature propagation and reduction in the vanishing-gradient problem, by enhancing it with multiple atrous convolution layers at different dilation rates. This setup enables the model to capture detailed semantic information without a substantial increase in computational complexity, making it suitable for real-time applications such as autonomous driving.

4.1.1 Atrous Spatial Pyramid Pooling

Atrous Spatial Pyramid Pooling (ASPP) is a technique designed to capture multi-scale information by applying atrous convolution with different dilation rates in parallel. This allows the network to effectively enlarge the field of view without losing resolution, which is crucial for semantic segmentation tasks. In ASPP, multiple parallel atrous convolutions with varying rates sample the input features at different scales, concatenating their outputs to form a final feature map that is rich in multi-scale context information. This approach helps the network to handle objects at various scales more efficiently.

Atrous Convolution Atrous convolution, also known as dilated convolution, is used to increase the receptive field of convolutional layers without losing spatial resolution. This is achieved by introducing zeros (or "holes") between the filter weights, effectively spreading the weights over a larger area. The dilation rate d controls the spacing between the weights. For example, a 3x3 filter with a dilation rate of 2 will have the same number of parameters as a standard 3x3 filter but will cover a 5x5 area. The general formula for atrous convolution is:

$$y[i] = \sum_{k=1}^{K} x[i + d \cdot k] \cdot w[k]$$

where d is the dilation rate, $w[k]$ are the filter weights, and K is the filter size.

Spatial Pyramid Pooling Spatial Pyramid Pooling (SPP) is a technique that pools features from different spatial bins or scales, allowing the network to consider multiple levels of context. In the context of atrous convolutions, this involves applying convolutions with different dilation rates in parallel to capture information at various scales. By doing so, SPP can create a rich, multi-scale representation of the input features, which is particularly useful for tasks like semantic segmentation where objects can vary significantly in size and shape.

Formulation of ASPP Formally, ASPP can be expressed as follows:

$$y = H_{3,6}(x) + H_{3,12}(x) + H_{3,18}(x) + H_{3,24}(x)$$

where $H_{K,d}(x)$ represents an atrous convolution with kernel size K and dilation rate d, applied to the input x.

DenseASPP enhances this approach by arranging atrous convolutional layers in a cascading manner, where the dilation rate increases progressively with each layer. The output of each layer is concatenated with the input feature map and the outputs of all preceding layers, leading to a dense feature pyramid that effectively captures semantic information at multiple scales.

4.2. Gated-SCNN

Gated-Shape CNN (Gated-SCNN), described in [6], introduces an innovative dual-stream architecture (Figure 3) where one stream processes traditional visual features while the other focuses exclusively on capturing shape attributes. The gating mechanism within the shape stream allows the model to selectively enhance features relevant for accurate boundary detection. This approach improves the precision of segmentation near object boundaries and enhances the model's ability to differentiate between closely situated objects, which is critical in cluttered urban environments.

4.2.1 Shape Stream and Gated Convolutional Layer

The shape stream in Gated-SCNN is dedicated to processing boundary-related information. It consists of residual blocks interleaved with Gated Convolutional Layers (GCL). The GCL uses higher-level activations from the regular stream to gate lower-level activations in the shape stream,

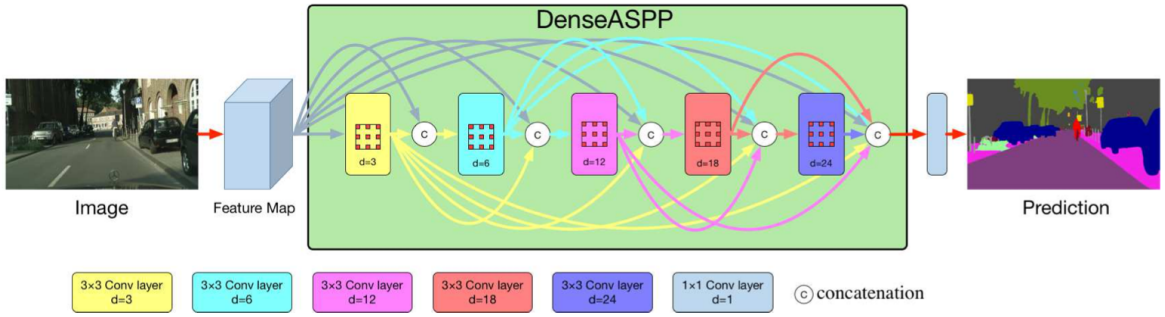

Figure 2. **DenseASPP architecture.** The figure illustrate DenseASPP in detail, the output of each dilated convolutional layer is concatenated with input feature map, and then feed into the next dilated layer. Each path of DenseASPP compose a feature representation of correspond scale.

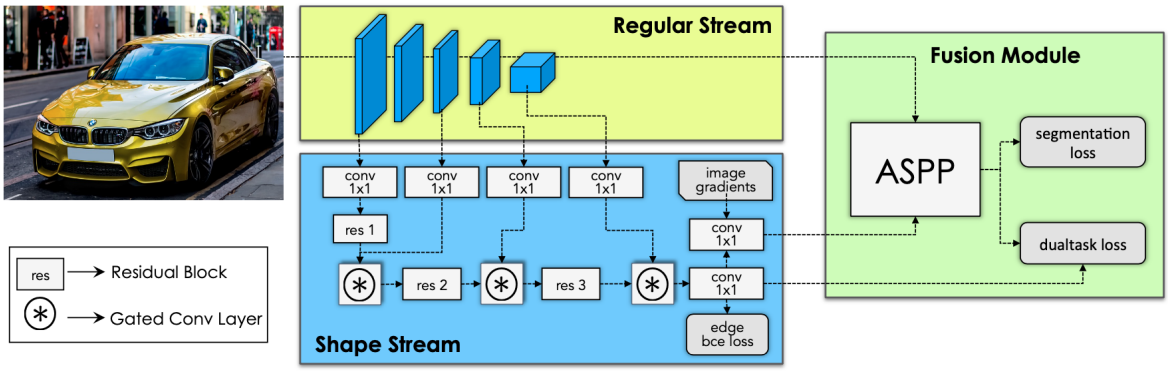

Figure 3. **Gated-SCNN architecture.** This architecture constitutes of two main streams: Regular stream and Shape stream. The Regular stream can be any backbone architecture. The shape stream focuses on shape processing through a set of residual blocks, Gated Convolutional Layers (GCL) and supervision. A Fusion module later combines information from the two streams in a multi-scale fashion using an Atrous Spatial Pyramid Pooling module (ASPP). High quality boundaries on the segmentation masks are ensured through a Dual Task Regularizer .

effectively removing noise and allowing the shape stream to focus on relevant boundary information. The attention map generated by concatenating feature maps from both streams undergoes a 1x1 convolution followed by a sigmoid function to create a gating mechanism. This mechanism enables the shape stream to adopt a shallow architecture that processes high-resolution images effectively, leading to sharper boundary predictions.

4.2.2 Loss Functions

Gated-SCNN employs four types of loss functions to enhance model performance:

- **Boundary Loss**: Uses binary cross-entropy (BCE) on predicted boundary maps, supervising the shape stream to ensure high-quality boundary detection.
- **Semantic Segmentation Loss**: Standard cross-entropy (CE) loss on predicted segmentation maps, updating parameters across both streams.
- **Dual Task Regularizer - Boundary to Segmentation (Forward)**: Ensures predicted boundaries align with ground-truth boundaries, promoting consistency between segmentation and boundary predictions.
- **Dual Task Regularizer - Segmentation to Boundary (Backward)**: Uses boundary predictions to refine semantic segmentation, enhancing alignment between predicted boundaries and semantic classes.

These loss functions work in tandem to improve both boundary and region accuracy, ensuring the Gated-SCNN performs effectively in real-world scenarios, particularly in urban environments.

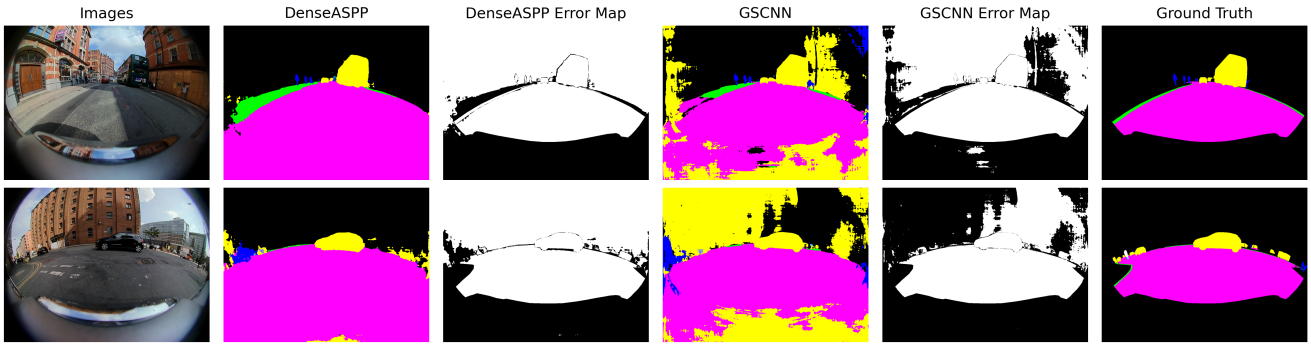

Figure 4. Qualitative results before fine-tuning

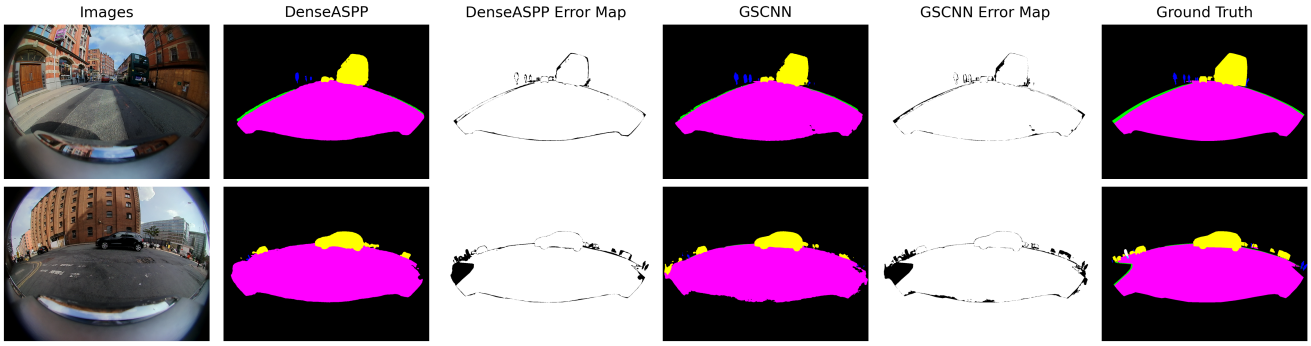

Figure 5. Qualitative results after fine-tuning

5. Experiments

5.1. Experiment 1: Pre-trained Model Evaluation

First, we evaluated the two pretrained models on the WoodScape dataset. Since their training dataset, Cityscapes, includes 19 semantic classes whereas WoodScape contains only 9, we created a mapping to associate each of the Cityscapes classes with a corresponding class in WoodScape. This mapping allowed us to test the pre-trained models on the new dataset. The initial results were unsatisfying, as illustrated by the poor quality of the predictions shown in figure 4. This outcome highlights the inadequacy of the models when directly applied to WoodScape, primarily due to the radial distortion in the images. Despite the semantic similarities between the classes in both datasets, the models struggled to generalize effectively. This experiment underscores the need to adapt models specifically for the WoodScape dataset using fine-tuning techniques. In the subsequent experiment, we applied these methods to enhance performance of the two pre-trained models.

5.2. Experiment 2: Fine-tuning

In the second experiment, we focused on retraining the two models to better adapt to the WoodScape dataset. Our objective was to evaluate if the models could learn the specific characteristics of WoodScape, given that the models pre-trained on CityScape did not perform adequately, as shown by the results of the first experiment. To achieve this, we adopted a fine-tuning strategy, which involves freezing the initial layers and retraining the remaining ones on the new examples. Specifically, for DenseASPP, we opted to freeze the first two Dense Blocks of the DenseNet-161 architecture serving as the backbone, while for Gated-SCNN, we decided to freeze the first four modules of the WideResNet backbone architecture.

5.2.1 Experimental Setup

The experiments were conducted on Google Colab with an 8GB GPU. Our training protocol followed the approach used for training models on the Cityscapes dataset.

For DenseASPP, we used the Adam optimizer with an initial learning rate of 0.0003 and a weight decay of 0.00001. The learning rate schedule involved multiplying the initial learning rate by $(1 - \frac{\text{current epoch}}{\text{max epochs}})^{0.9}$. Instead, for GSCNN, we used SGD with an initial learning rate of 0.0005 and a weight decay of 0.00001, following the same learning rate scheduler as DenseASPP. Both models were trained for 5 epochs with a mini-batch size equal to 3.

5.2.2 Evaluation Metrics

Performance was measured using standard metrics such as Mean Intersection over Union (mIoU) and F1-score.

In particular, Mean Intersection over Union is highly useful in assessing the accuracy of segmentation results by measuring the overlap between predicted and ground truth segmentation masks, and it is defined as:

$$MeanIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i$$

where, for each i, the Intersection over Union is calculated as the intersection of the predicted and ground truth masks divided by their union:

$$IoU = \frac{\text{Area of Intersection}}{\text{Area of Union}}$$

These metrics provide a comprehensive view of model performance across various classes and scene complexities.

5.2.3 Results

The results in Figure 5, show that both models, after fine-tuning, effectively captured the radial distortion in the WoodScapes dataset images, learning to correctly recognize the various semantic classes within the images, demonstrating greater accuracy and robustness in identifying and categorizing different elements in the dataset.

As shown in the Table 2, the Gated-SCNN model outperformed DenseASPP across both evaluation metrics considered. Additionally, according to the results obtained in Table 1, Gated-SCNN exhibited superior boundary precision and class differentiation, attributed to the inclusion of a shape stream that significantly enhanced edge predictions. (These evaluations were computed on the full images without cropping)

Class	DenseASPP	Gated-SCNN
void	**95.16%**	94.07%
road	**91.88%**	89.10%
curb	43.35%	**44.65%**
person	29.90%	**40.11%**
rider	0.00%	**40.73%**
vehicles	69.67%	**71.11%**
bicycle	8.73%	**36.14%**
motorcycle	1.55%	**32.55%**
traffic_sign	**6.59%**	4.10%

Table 1. Comparison of DenseASPP and Gated-SCNN results across different classes.

Model	mIoU	f1-score
DenseASPP	38.54%	0.46
Gated-SCNN	50.28%	0.62

Table 2. Comparison of mIoU and f1-score for DenseASPP and Gated-SCNN models.

6. Conclusions

In this study, we explored the capabilities of two state-of-the-art semantic segmentation models, DenseASPP and Gated-SCNN, in handling the unique challenges presented by the WoodScapes dataset, characterized by fisheye camera's images with radial distortion. Our findings can be summarized as follows:

- **Effectiveness of Fine Tuning**: By fine-tuning the models on the WoodScapes dataset, we significantly improved their performance compared to their initial application. This highlights the importance of domain-specific adaptation for models trained on standard datasets like Cityscapes.
- **Model Performance**: Although both pre-trained models achieved good results after the adaptations, the Gated-SCNN model consistently outperformed DenseASPP, achieving higher mIoU and F1-scores. Specifically, Gated-SCNN's shape stream and gating mechanism proved crucial in enhancing boundary precision and class differentiation, which are vital for the complex scenes found in automotive images.
- **Architectural Insights**: Our analysis underscores the significance of advanced architectural elements such as Atrous Spatial Pyramid Pooling (ASPP) and Gated Convolutional Layers (GCL). These components play a pivotal role in capturing multi-scale context and refining feature extraction, thereby improving overall segmentation quality.
- **Application to Real-World Scenarios**: The results demonstrate the robustness of Gated-SCNN in handling distorted images, making it a promising approach for real-world applications in autonomous driving where fisheye cameras are commonly used.

Overall, this study contributes to the understanding of how different semantic segmentation architectures perform in challenging imaging conditions. Future work could explore further enhancements, such as incorporating additional data augmentation techniques and experimenting with other advanced models to continue improving segmentation performance in distorted image scenarios.

For the complete code and detailed implementation of the experiments, please refer to our GitHub repository at https://github.com/andrea3425.

References

[1] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S. Lew. A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, 7(2):87–93, 2018.

[2] Shijie Hao, Yuan Zhou, and Yanrong Guo. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406(1):302–321, 2020.

[3] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1(1):2261–2269, 2017.

[4] Yujian Mo, Yan Wu, Xinneng Yang, Feilin Liu, and Yujun Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493(1):626–646, 2022.

[5] Uroosa Sehar and Muhammad Luqman Naseem. How deep learning is empowering semantic segmentation. *Multimedia Tools and Applications*, 81(1):30519–30544, 2022.

[6] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. *IEEE International Conference on Computer Vision*, 1(1):5229–5238, 2019.

[7] Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126(1):106669, 2023.

[8] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 1(1):2843–2851, 2018.

[9] Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sumanth Chennupati, Sanjaya Nayak, Saquib Mansoor, Xavier Perrotton, and Patrick Perez. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. *arXiv preprint arXiv:1905.01489*, 1(1):1–11, 2021.