공학석사 학위논문

실시간 의미분할을 위한 효율적인 어텐션 메카니즘과 원근변환

Efficient Attention Mechanism and Perspective Transform in Realtime Semantic Segmentation

울산대학교 대학원

전기전자컴퓨터 공학과

정창현

실시간 의미분할을 위한 효율적인 어텐션 메카니즘과 원근변환

지도교수 조강현

이 논문을 공학석사학위 논문으로 제출함

2023 년 7월

울산대학교 대학원

전기전자컴퓨터공학과

정창현

This certifies that dissertation of Changhyun Jeong

Committee Chair

Prof. Heejun Kang

강 희준

Committee Member and Supervisor

Prof. Kanghyun Jo

Committee Member

Prof. Youngsoo Suh

Department of Electrical, Electrical and Computer Engineering

University of Ulsan, South Korea

May 2023

MASTER OF SCIENCE

Efficient Attention Mechanism and Perspective Transform in Realtime Semantic Segmentation

The Graduate School

of the University of Ulsan

Department of

Electrical, Electronic and Computer Engineering

Jeong, Chang-hyun

Efficient Attention Mechanism and Perspective Transform in Realtime Semantic Segmentation

Supervisor : Jo, Kang-hyun

A Dissertation

Submitted to

the Graduate School of the University of Ulsan

In partial Fulfillment of the Requirements

for the Degree of

Masters of Science

by

Jeong, Chagn-hyun

Department of Electrical

Electronic and Computer Engineering

Ulsan, Korea

July 2023

ACKNOWLEDGEMENT

I would like to express my profound gratitude to my advisor, Professor Kanghyun Jo, for his unwavering support and guidance throughout my research journey. I am deeply grateful to my committee members professor Kang and professor Seo for their invaluable insights and constructive feedback.

Additionally, I extend my heartfelt thanks to my colleagues in the ISLAB for their camaraderie and support. Special appreciation goes to our captain, Yulkyong Lee, the ever-diligent Jehwan Choi, and the warm-hearted Junmyong Kim. Though the journey was challenging, it was a truly meaningful and transformative time in my life.

[ABSTRACT]

Efficient Attention Mechanism and Perspective Transform in Realtime Semantic Segmentation

Changhyun Jeong

School of Electrical Engineering,

The Graduate School of the University of Ulsan

Supervised by Prof. Kanghyun Jo

Pedestrian guidance systems are essential for assisting individuals with cognitive impairments and enhancing urban mobility. In this paper, we propose an efficient attention mechanism and perspective transform for real-time semantic segmentation to improve the performance of pedestrian guidance systems. We utilized the STDC2 semantic segmentation machine learning model as our base model and integrated the CBAM and to enhance its efficiency and accuracy.

Although the addition of these attention modules led to increased accuracy, the model's accuracy-computation tradeoff was not significantly improved. To address this challenge, we investigated the use of the Korean pedestrian dataset for training our model. However, the dataset's viewpoint is downward facing, which is not suitable for a pedestrian guidance system. We employed perspective transformation techniques to adapt the dataset for training a model that can effectively guide pedestrians in urban environments.

The proposed combination of an efficient attention mechanism and perspective transform allows for the development of a robust and accurate real-time semantic segmentation model. Our approach aims to enhance the performance of pedestrian guidance systems, providing better assistance to individuals with cognitive impairments and improving overall urban mobility.

Index

List of Figures

List of Tables

1. Introduction

1.1 Pedestrian Guidance System

The field of machine learning has revolutionized many industries and has the potential to greatly impact the lives of people in a positive way. In particular, the use of machine learning in image processing has opened up many new possibilities for various applications, including pedestrian guidance systems. These systems have the potential to greatly benefit the elderly and individuals with dementia by providing them with assistance and guidance as they navigate their environment.

Dementia and Alzheimer's Disease: Dementia is an umbrella term for a group of neurological disorders characterized by a progressive decline in cognitive abilities, including memory, thinking, and problem-solving skills, which ultimately impacts an individual's ability to perform daily activities. Alzheimer's disease is the most prevalent form of dementia and is a progressive neurodegenerative disorder that primarily affects older adults. The disease is marked by the accumulation of abnormal protein deposits in the brain, leading to the degeneration of brain cells and cognitive decline.

As of 2021, it was estimated that approximately 50 million people worldwide suffered from dementia, and this number is projected to increase to 152 million by 2050, primarily due to the aging global population. Alzheimer's disease accounts for 60-80% of dementia cases. In the United States alone, over 6 million people were living with Alzheimer's disease, with a new diagnosis occurring every 65 seconds.

Navigating the environment can be particularly challenging for individuals with dementia and Alzheimer's disease. Some of the reasons for these difficulties include:

Memory Loss: As dementia progresses, individuals often experience short-term memory loss, which makes it difficult to remember directions, landmarks, or even familiar locations. This can lead to disorientation and an inability to navigate independently.

Dementia and Alzheimer's disease can affect an individual's spatial awareness, making it challenging to perceive distances and the relative positions of objects in the environment. This can hinder their ability to navigate complex routes or avoid obstacles. Individuals with dementia may struggle with problem-solving and decision-making, making it difficult for them to determine the best route to their destination or adapt to changes in the environment, such as construction or road closures.

Dementia and Alzheimer's disease can cause disorientation, even in familiar surroundings, leading to confusion and anxiety. This can further exacerbate navigational difficulties and increase the likelihood of wandering incidents. Given the navigational challenges faced by individuals with dementia and Alzheimer's disease, there is a pressing need for a pedestrian guidance system specifically designed for this population. Such a system can:

Provide tailored assistance: A pedestrian guidance system can offer personalized navigation support, taking into account the unique needs and abilities of individuals with dementia and Alzheimer's disease, such as simplified routes and visual or auditory cues.

Promote independence: By providing real-time guidance, the system can help individuals with dementia maintain their independence and engage in daily activities with reduced reliance on caregivers, ultimately improving their quality of life.

Enhance safety: A pedestrian guidance system can help prevent wandering incidents and ensure the safety of individuals with dementia and Alzheimer's disease by guiding them along safe routes and alerting caregivers in case of deviations or emergencies.

Alleviate caregiver burden: By assisting with navigation, the pedestrian guidance system can reduce the emotional and physical strain on caregivers, allowing them to focus on other aspects of care and support.

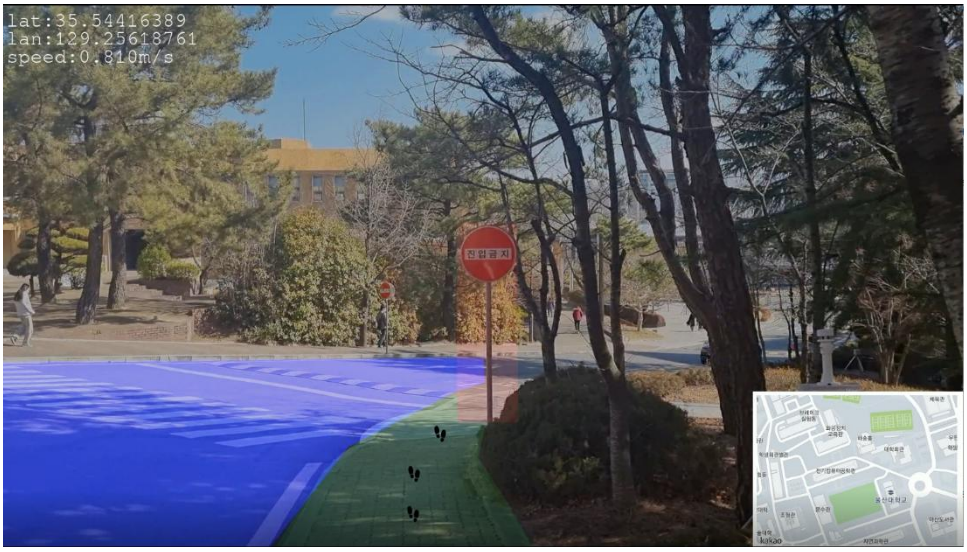


Fig 1. Example screen of the Pedestrian Guidance System

1.2 Research Content

In this research paper, we will explore various aspects of developing an efficient attention mechanism for pedestrian guidance systems tailored to assist individuals with dementia and Alzheimer's disease. The following sections outline the key areas of investigation:

Datasets for Pedestrian Guidance System: We will examine two primary datasets that can be utilized for developing a pedestrian guidance system:

a. CityScapes[1]: The CityScapes dataset is a large-scale, diverse dataset comprising high-quality pixel-level annotations of urban street scenes from 50 cities across different countries, captured during various weather conditions and seasons. This dataset offers a comprehensive representation of real-world urban environments, which can be valuable for training and evaluating our pedestrian guidance system.

b. Korean AIHub Pedestrian Dataset[2]: This dataset contains annotated pedestrian images captured in various urban environments in South Korea. The dataset offers a diverse set of images featuring pedestrians in different poses, clothing, and occlusions, making it suitable for training and evaluating pedestrian detection and guidance models.

The paper will compare the pros and cons of each dataset in terms of diversity, annotation quality, and suitability for our pedestrian guidance system, highlighting the advantages and limitations of each dataset.

Data Augmentation in Image Datasets: Data augmentation is a crucial technique for improving the performance of deep learning models by generating additional training data from the existing dataset. We will investigate various data augmentation techniques suitable for image datasets, such as rotation, scaling, flipping, and cropping. The research will also explore advanced augmentation techniques like mixup, cutout, and cutmix to enhance the robustness and generalizability of our pedestrian guidance model.

Efficient Attention Mechanism: Attention mechanisms have been shown to improve the performance of deep learning models by enabling them to focus on the most relevant features in the input data. In this research paper, we will study various efficient attention mechanisms, such as the Squeeze-and-Excitation[3] (SE) module, the Efficient Channel Attention (ECA) module, and the Convolutional Block Attention Module[4] module. We will analyze their computational efficiency, memory requirements, and potential for integration into our pedestrian guidance system.

Real-time Semantic Segmentation Models: Semantic segmentation is an essential component of a pedestrian guidance system, as it enables the model to understand and interpret the surrounding environment. We will review state-of-the-art real-time semantic segmentation models, such as STDC[5] and BiSeNet[6] and analyze their performance, computational complexity, and suitability for deployment on edge devices.

1.3 Research Target

In this research paper, we aim to develop an efficient attention mechanism for a pedestrian guidance system designed to assist individuals with dementia and Alzheimer's disease. Our research targets are as follows:

Dataset Application and Comparison: We will apply the CityScapes dataset and the Korean AIHub Pedestrian dataset to train and evaluate the pedestrian guidance system. We will compare the results obtained using each dataset to assess their impact on the model's performance. This comparison will provide insights into the advantages and limitations of each dataset and guide our choice of the most appropriate dataset for further development of the guidance system.

Attention Module Experiments: We will experiment with the application of different attention modules, such as the SE module and ECA module on state-of-the-art attention models. These experiments will help us determine the most efficient and effective attention mechanism for our pedestrian guidance system. We will analyze the performance, computational complexity, and memory requirements of each attention module to ensure suitability for deployment on edge devices.

Model Optimization for Accuracy/Speed Tradeoff: The final research target is to optimize the selected model for a balance between accuracy and speed. We will explore various techniques, including model pruning, quantization, and knowledge distillation, to reduce the computational complexity and memory footprint of the model without sacrificing its performance. The optimized model will allow us to develop a pedestrian guidance system that provides accurate and real-time assistance to individuals with dementia and Alzheimer's disease while adhering to the constraints of edge devices.

By achieving these research targets, we aim to contribute to the development of an efficient and effective pedestrian guidance system that can enhance the safety, independence, and overall quality of life for individuals with dementia and Alzheimer's disease.

2. Literature Review

2.1 Attention Mechanisms in computer vision

The attention mechanism is an influential technique in the realm of modern neural networks, especially in computer vision. Its objective aligns with the goal of computer vision, to emulate human visual perception through algorithmic structures.

Let's consider human visual capabilities, which have directly inspired attention mechanisms. In human vision, light from an object, for instance, a leaf, reaches the macula - the primary functional region of the retina. But, when there are multiple objects in view, focusing on a single object requires the human attention system to apply a range of filters, creating a blurred or "bokeh" effect around the central object, thereby focusing attention on it.

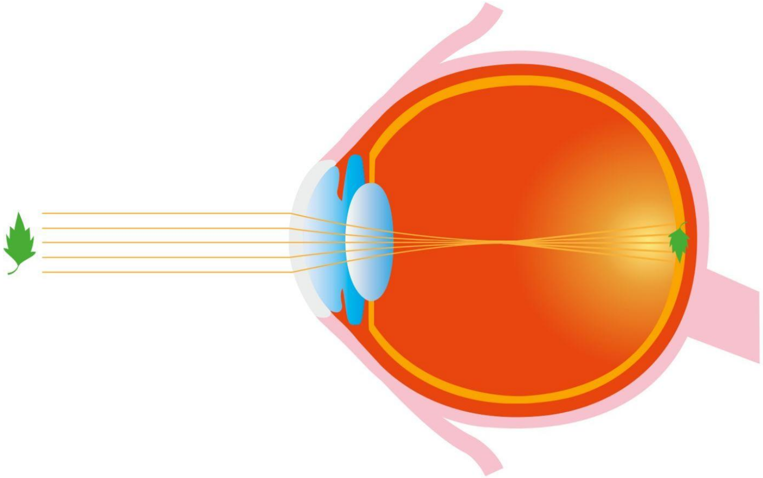


Fig 2. Illustration of human vision focus.

In the diagram above, the light traverses from the object of interest (which in this case is a leaf) to the Macula, which is the primary functional region of the Retina inside the eye. However, what happens when we have multiple objects in the field of view?

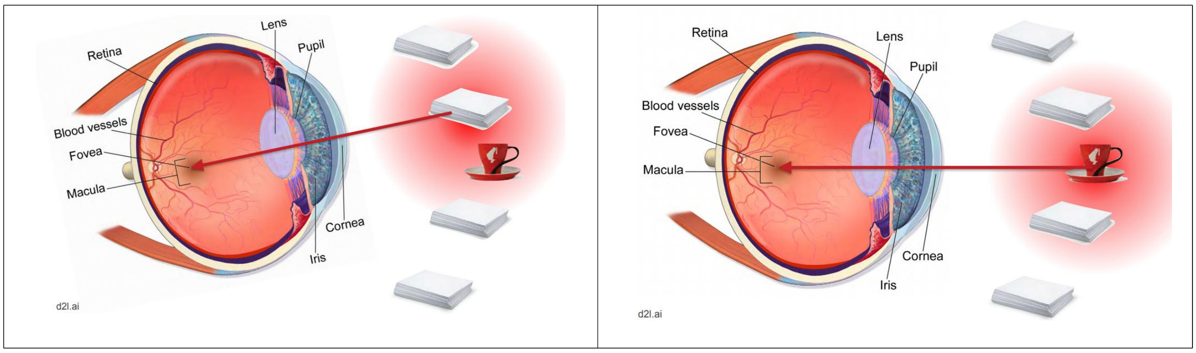


Fig 3. Illustration of attention changes in human vision

When we must focus on a single object when there is an array of diverse objects in our field of view, the attention mechanism within our visual perception system uses a sophisticated group of filters to create a blurring effect (similar to that of "Bokeh" in digital photography) so the object of interest is in focus, while the surrounding is faded or blurred.

The concept of Attention Mechanisms gained popularity in the field of Natural Language Processing (NLP), as elaborated in "Attention Is All You Need"[10]. The paper proposed that attention could be computed using three core components: Query, Key, and Value. This mechanism was later adapted for computer vision in the SAGAN[11] paper, which introduced the Self Attention Module, represented diagrammatically as f(x), g(x), and h(x) for query, key, and value respectively.

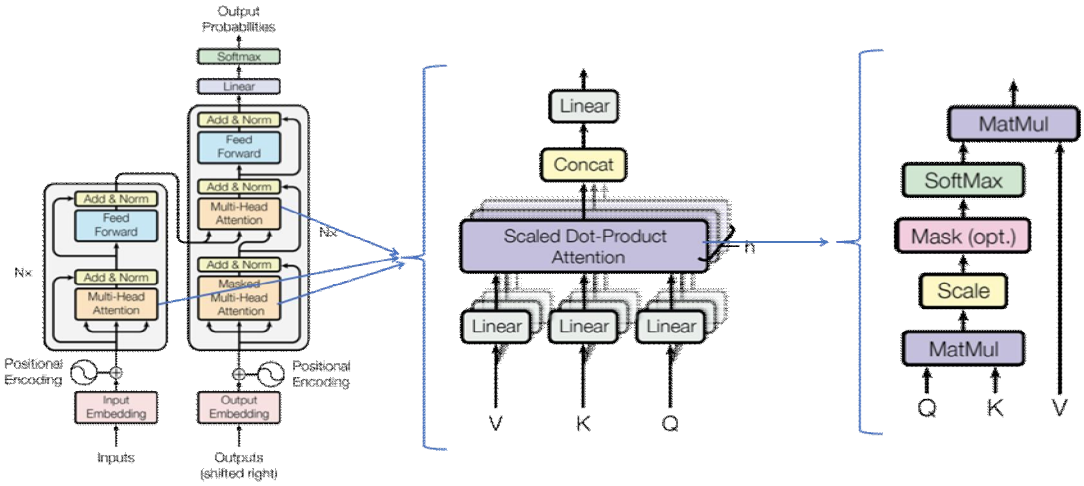


Fig 4. Transformer Architecture, Scaled Dot Product Attention, and Multi-Head Attention.

2.2 Squeeze-and-Excitation Networks

Squeeze-and-Excitation Networks (SENet), introduced by Jie Hu, Li Shen, and Gang Sun in 2018, have made a significant contribution to the attention mechanism's evolution. SENet fundamentally introduces a novel architectural unit known as the "SE block". This block imbues the model with the capacity to adaptively recalibrate channel-wise feature responses.

It's important to note that SENet isn't a standalone architecture. Instead, it operates as a supplemental unit that enhances existing architectures such as ResNet and Inception by incorporating a context-aware ability. This augmentation improves the base architecture's performance by allowing it to attend to the most salient features in the input data.

The SE block comprises two key stages:

The Squeeze operation: This operation's primary aim is to capture the global information present in the input feature maps. Accomplished using Global Average Pooling (GAP), the Squeeze operation reduces the spatial dimensions (height and width) to 1, resulting in a tensor of shape (batch size, channels, 1, 1). This operation essentially condenses the spatial information, enabling the capture of channel-wise statistics.

The Excitation operation: This stage is designed to capture channel-wise dependencies in their entirety. Structured as a two-layer fully connected (FC) network, the Excitation operation employs a sigmoid activation function at its output stage to ensure that output values fall within the range of 0 and 1. These output values can be interpreted as channel-wise weights that are used to recalibrate the original feature maps.

The SE block's unique feature lies in its channel-wise approach, where it derives parameters for each channel separately based on globally averaged pooled information. This approach allows it to learn and explicitly model interdependencies between channels. Consequently, the SE block can highlight informative features and suppress less relevant ones, thereby enhancing the model's focus on crucial information.

The Squeeze-and-Excitation Networks' effectiveness has been demonstrated through their application in various deep learning architectures, resulting in substantial performance improvement. This powerful component, with its operational efficiency and simplicity, has been integrated into a wide array of existing deep learning models, reinforcing their capacity to focus on the most salient features in input data.

2.3 Convolutional Block Attention Modules

The Convolutional Block Attention Module (CBAM) is a pivotal aspect of computer vision. It was first introduced in 2018, but the fundamental concept had been presented earlier in SCA-CNN[8]. The latter demonstrated the effectiveness of using Spatial Attention and Channel Attention in conjunction - the two essential elements of CBAM - for image captioning tasks. CBAM was the first to illustrate the module's broad applicability, particularly for image classification and object detection tasks.

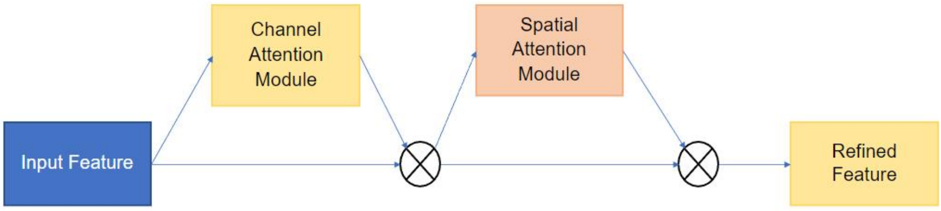


Fig 5. Convolutional Block Attention Module layout

CBAM comprises two sequential sub-modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM), applied in that order. In any convolutional layer, both the input and the output are tensors characterized by three-dimensional metrics: height (h), width (w), and the number of channels (c) or depth of the tensor, represented by the total number of feature maps.

The term "spatial" pertains to the space within each feature map. Spatial attention signifies the attention mask on the feature map, which essentially enhances the features defining the object of interest. By refining the feature maps through Spatial Attention, the input to the subsequent convolutional layers is improved, thereby enhancing the overall model's performance.

Channel attention, on the other hand, provides weights for each channel, thereby magnifying channels that contribute most to the learning process and overall model performance. It's important to note that, even though many feature maps (slices of the tensor) might seem identical due to small weights (close to zero) in convolutional layers, they play a crucial role in learning different features. Some filters are designed for learning horizontal and vertical edges, while others learn particular textures in the image. This observation was a critical influence for "GhostNet: More Features from Cheap Operations"[12].

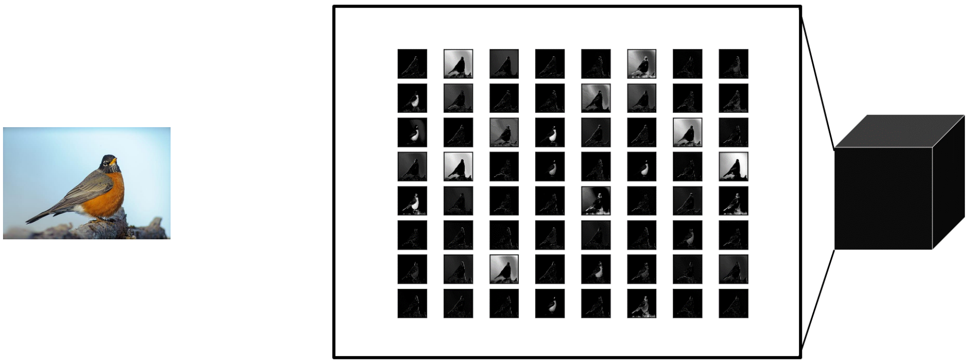


Fig 6. Feature Maps representation as a Tensor[9]

Is there a need to utilize both, or would one be sufficient? Well, the answer is somewhat nuanced; technically, both yes and no. The authors, in their code implementation, offer the option to use only the Channel Attention and disable the Spatial Attention. However, to obtain optimal results, it is recommended to employ both. In simpler terms, the Channel Attention module discerns which feature map is critical for learning and enhances or "refines" it accordingly. Conversely, Spatial Attention indicates what parts within the feature map are essential to learn. Together, they enhance the Feature Maps more robustly, substantiating the significant improvement in model performance.

2.3.1. Spatial Attention Module (SAM)

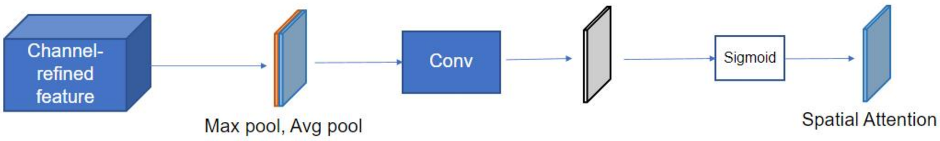


Fig 7. Spatial Attention Module

The Spatial Attention Module (SAM) consists of a three-step sequential operation. The first segment is referred to as the Channel Pool, where the Input Tensor with dimensions (c × h × w) is broken down into 2 channels, i.e. (2 × h × w). Each of these two channels represents Max

Pooling and Average Pooling across channels. This serves as the input for the convolution layer that outputs a 1-channel feature map, i.e., the output dimension is (1 × h × w). This convolution layer preserves spatial dimensions and uses padding to do so. In code, the convolution is followed by a Batch Norm layer that normalizes and scales the convolution output. Although there is an option to use a ReLU activation function post the Convolution layer, by default, it uses only Convolution + Batch Norm. The output then passes through a Sigmoid Activation layer. Being a probabilistic activation, Sigmoid will map all the values to a range between 0 and 1. This Spatial Attention mask is then applied to all feature maps in the input tensor using a simple element-wise product. The authors validated various approaches of computing Spatial Attention using SAM in the ImageNet classification task, as shown in the results from the paper.

2.3.2. Channel Attention Module (CAM)

The Channel Attention Module (CAM) operates sequentially like the SAM but is more complex. Upon first observation, CAM resembles the SE layer.

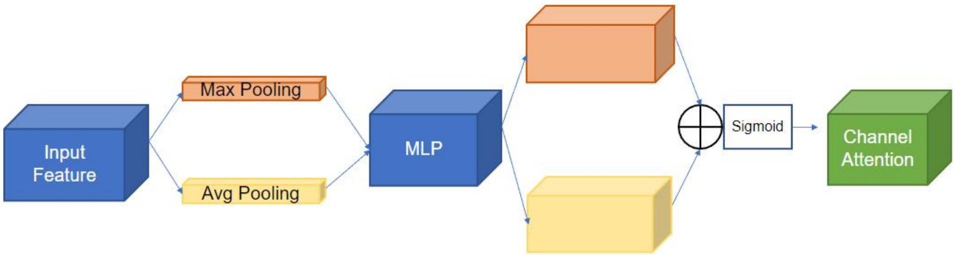


Fig 8. Channel Attention Module

o what distinguishes the Squeeze Excitation from the Channel Attention Module?

Before diving into this, let's quickly recap the Squeeze Excitation Module. It comprises components like Global Average Pooling (GAP) and a Multi-layer Perceptron (MLP) network, which is influenced by a reduction ratio (r) and sigmoid activation. The input to the SE block is a tensor of dimension (c × h × w). GAP is essentially an Average Pooling operation where each feature map is reduced to a single pixel, so each channel is now decomposed to a (1 × 1) spatial dimension. Therefore, the output dimension of the GAP is a 1-D vector of length c, represented as (c × 1 × 1). This vector is then input to the MLP network, which has a bottleneck whose width or number of neurons is decided by the reduction ratio (r). The output vector from this MLP then goes through a sigmoid activation layer, which maps the values in the vector within the range of 0 and 1.

The CAM is quite similar to the Squeeze Excitation layer but with a minor adjustment. Instead of reducing the Feature Maps to a single pixel by GAP, it breaks the input tensor into two subsequent vectors of dimensionality (c × 1 × 1). One of these vectors is generated by GAP,

2.4 Perspective Transformation in Image Processing

Perspective transformation, also known as homography, plays an essential role in many image processing and computer vision applications, including image stitching, panorama generation, image rectification, and the simulation of camera motion. It enables the transformation of images to cater to various viewing angles, offering a more comprehensive understanding and utilization of the available data.

In the context of computer vision and machine learning, it's common to encounter datasets that may not perfectly align with the desired task. For instance, images taken from different viewpoints or angles could negatively impact the performance of models, especially those trained for tasks like object detection or semantic segmentation. When such datasets are employed in training models, the resultant models may exhibit reduced effectiveness when applied to normal human pedestrian viewed images due to the disparity in perspective.

To address such issues, perspective transformation techniques have been utilized to adapt the dataset to fit the required task. Through the application of a homography matrix, images can be transformed to simulate a change in camera perspective. This matrix defines the transformation from one plane to another, enabling the conversion of an image taken from a bird's-eye view (downwards-looking) to one that appears to have been captured from a horizontal perspective.

Several methods have been proposed to estimate this homography matrix, typically involving point correspondences between the two images. Popular methods for homography estimation include Direct Linear Transform (DLT), Normalized Direct Linear Transform (NDLT), and the Random Sample Consensus (RANSAC) algorithm for outlier rejection.

The use of perspective transformation to augment the existing dataset has several benefits. By providing a different viewpoint, the transformed dataset can effectively increase the diversity of the training data, thereby improving the robustness and generalizability of the model. Furthermore, the transformation process can help adjust images to fit the model's expectation of the input, thus enhancing its performance in real-world scenarios.

While perspective transformation presents a promising avenue for the utilization of misaligned datasets, certain challenges may arise, such as loss of image information due to the transformation, the introduction of empty or undefined regions in the resultant images, and the need for careful parameter tuning to ensure an optimal transformation. Despite these challenges, the adoption of perspective transformation techniques, when combined with other data augmentation methods, could potentially unlock the full potential of existing datasets for a wide array of computer vision tasks.

3. Real Time Semantic Segmnetation

Advancements in real-time semantic segmentation have been fundamental in enhancing numerous applications such as autonomous driving, video surveillance, and medical imaging. Despite the progress made, the challenge lies in striking a balance between model speed and prediction accuracy.

3.1 Speed-Accuracy Trade-offs in Existing Models

Several works have proposed various strategies to improve real-time semantic segmentation. For instance, the ICNet[13] and the 'Real-time Image Segmentation via Spatial Sparsity'[14] focused on creating a practical and fast semantic segmentation system with a reasonable prediction accuracy. They prioritized improving the inference speed without a significant compromise on the quality of the segmentation.

One approach used in these studies involves limiting the input size to decrease computational complexity. However, this method—achieved through cropping or resizing—often results in the loss of crucial spatial details, particularly around boundaries, thereby diminishing the accuracy of predictions.

Other works, like Xception[15], employ channel pruning as a speed optimization technique. This approach reduces the width of the feature map, resulting in a smaller and faster network. While this method is efficient on both CPU and GPU without necessitating any special implementation, the information loss from pruning can sometimes be challenging to recover, causing a decrease in segmentation accuracy.

The ENet (Efficient Neural Network)[16] adopts a different approach by dropping the downsampling operations in the final stage. Instead, it employs upsampling operations, which inversely affect the discriminative ability of the model. Despite the larger receptive field allowed by operating filters on downsampled images, this approach encounters drawbacks related to the loss of spatial information and the necessity for full-pixel segmentation.

To mitigate the loss of spatial detail caused by the above strategies, some researchers have utilized a U-shape[17] structure. By fusing hierarchical features of the backbone network, this structure gradually improves spatial resolution, thereby recovering some lost details. Despite its advantages, this technique introduces two main issues: It reduces the model's speed due to additional computation, and most spatial information lost during pruning remains unrecoverable.

3.2 Bilateral Segmentation Network (BiSeNet)

The Bilateral Segmentation Network (BiSeNet) offers an innovative approach to real-time semantic segmentation by finding a balance between model speed and accuracy. This novel architecture introduces a dual-path system: the Spatial Path (SP) preserves spatial details with a small stride, while the Context Path (CP) quickly downsamples to capture larger context.

Two notable features are the Feature Fusion Module (FFM) and the Attention Refinement Module (ARM), designed for efficient feature combination and stage-by-stage feature refinement, respectively. ARM, without requiring upsampling, leverages global average pooling for context capture. This results in a system that performs well on Cityscapes, CamVid, and COCO-Stuff datasets, showing significant speed improvements without compromising on accuracy.

The Spatial Path is a three-layer convolutional sequence that keeps spatial resolution at 1/8 of the original image, capturing abundant spatial information thanks to the large size of feature maps.

The Context Path complements the Spatial Path by providing an expansive receptive field. Various methods such as pyramid pooling and large kernels can enlarge the receptive field, but these tend to be computationally heavy and slow. The Context Path provides a balance between a large receptive field and computational efficiency.

3.2.1 Attention Refinement Module

The ARM, a key component in BiSeNet V2, captures high-level semantic information from the input image. It refines output feature maps of certain layers within the Context Path using global context extraction and channel-wise attention for feature map refinement. If necessary, it can also fuse feature maps from multiple layers.

3.2.2 Feature Fusion Module

The FFM integrates and refines feature maps from the Context Path and Spatial Path. It upscales the Spatial Path feature maps to match the resolution of those from the Context Path. Then, they are combined and further refined using convolutions, batch normalization, and ReLU activations. A channel-wise attention mechanism helps emphasize important features and suppress less critical ones. The final output feature maps, which retain both rich context and fine-grained spatial details, are upsampled to the original image resolution to yield pixel-wise semantic segmentation. Overall, the FFM contributes to the effective balance of speed and accuracy in BiSeNet V2.

3.3 STDC2

The Short-Term Dense Concatenate network (STDC2) presents an innovative solution to some of the drawbacks associated with popular dual flow networks such as BiSeNet. The STDC2 model eliminates time-consuming procedures by simplifying the approach to encoding spatial information. This network also avoids the inefficiencies of pre-trained tasks, like image classification, that are often incorporated into image segmentation frameworks. By minimizing structural redundancy, the STDC2 model showcases a novel and efficient architecture for real-time segmentation tasks. This architecture includes the gradual reduction of feature map dimensions and their aggregation for image representation. In the decoder, a detail aggregation module allows for the integration of spatial information learning into the lower level in a single-stream way. Lastly, the fusion of low-level features with deep features produces the final segmentation prediction.

3.3.1 STDC Module

The STDC Module, a central part of the STDC2 architecture, plays a significant role in refining spatial-temporal information within input feature maps. This component allows for improved semantic segmentation performance by using a refined process involving multiple steps.

Feature Map Refinement: Feature maps from multi-scale branches are combined, followed by the application of several convolutional layers, Batch Normalization, and ReLU activation functions. This approach helps to integrate and refine information from different scales.

Channel-wise Attention: Similar to the Attention Refinement Module (ARM) in BiSeNet V2, the STDC Module uses a channel-wise attention mechanism. This mechanism emphasizes significant channels and suppresses less critical ones. The refined feature maps undergo global average pooling and are passed through a fully connected layer followed by a sigmoid activation function. This process generates channel-wise attention weights which are applied to the feature maps.

Output Feature Maps: The STDC Module enhances spatial-temporal information in the output feature maps. This enriched information is critical for precise pixel-wise semantic segmentation, particularly in video-based applications.

In summary, the STDC Module in STDC2 offers a significant contribution to the process of capturing and refining spatial-temporal information within input feature maps. By employing the STDC Module, STDC2 provides improved performance in semantic segmentation, particularly in real-time applications that demand high precision and efficiency. The adaptive processing of complex spatial-temporal patterns offered by the STDC Module makes it a valuable tool in the field of semantic segmentation, offering a robust solution for various applications.

4. Datasets

4.1 CityScapes

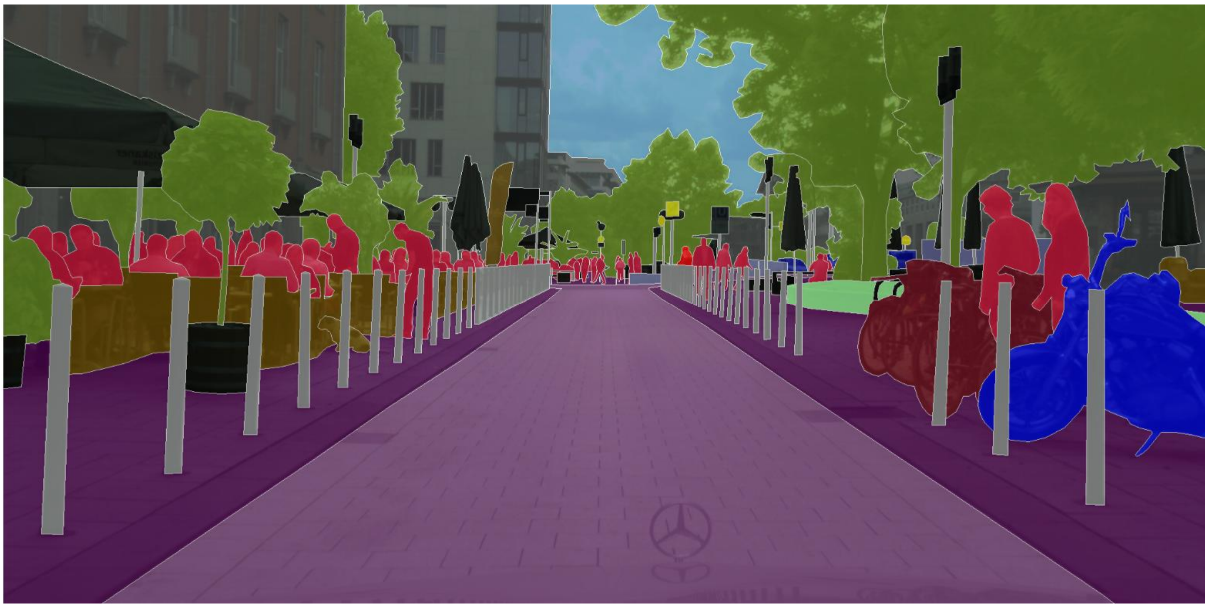


Fig 9. Example of CityScapes Dataset

The Cityscapes dataset is a large-scale dataset for semantic segmentation in urban street scenes. It was first introduced in 2016 by researchers from the University of Tubingen, Germany, and Daimler AG. Semantic segmentation is a computer vision task that involves classifying each pixel in an image into a specific category or label. In the case of the Cityscapes dataset, these categories correspond to various urban scene elements such as cars, pedestrians, buildings, roads, sidewalks, traffic signs, etc.

Cityscapes is specifically designed to help train and evaluate deep learning models for semantic segmentation in the context of autonomous driving and urban scene understanding. The dataset contains high-quality pixel-level annotations, which makes it suitable for developing and benchmarking state-of-the-art semantic segmentation algorithms.

The dataset consists of 5,000 high-quality annotated images, split into 2,975 for training, 500 for validation, and 1,525 for testing. These images have been collected from 50 different cities in Germany during different seasons and times of the day.

The images in the Cityscapes dataset are captured at a high resolution of 2048x1024 pixels, which provides detailed information for semantic segmentation algorithms to exploit.

The Cityscapes dataset provides two levels of annotation detail: fine and coarse. Fine annotations are available for the 5,000 images mentioned earlier, while an additional 20,000 images have coarse annotations. The fine annotations include pixel-level labels for 30 classes, with 19 of these classes being used for evaluation. Coarse annotations offer a lower level of detail, which can be useful for pretraining or experimenting with model architectures.

In addition to semantic segmentation, the Cityscapes dataset also provides instance-level annotations for certain object classes, such as cars, pedestrians, and bicycles. These annotations can be used for instance segmentation tasks, where the goal is to not only identify object classes but also to separate individual instances of those classes.

The standard evaluation metric used for the Cityscapes dataset is the mean Intersection-over-Union (mIoU). It measures the overlap between the predicted segmentation and the ground truth annotation, averaged across all classes. This metric is widely used in the semantic segmentation literature and allows for a fair comparison between different models and approaches.

Overall, the Cityscapes dataset has played a crucial role in advancing research in semantic segmentation and urban scene understanding. It has been widely adopted by the computer vision community and has led to numerous publications and advancements in the field.

The Cityscapes has greate categories, but problem is all the images are from europe, so road and sidewalks visual features are pretty different. So we cant use it on our korean envicronment.

Table 1. Categories of CityScapes dataset

Group	Classes
flat	road, sidewalk, parking, rail track
human	person, rider
vehicle	car, truck, bus, rails, motorcycle, bicycle, caravan, trailer
construction	building, wall, fence, guard rail, bridge, tunnel
object	pole, pole group, traffic sign, traffic light
nature	vegetation, terrain
sky	sky
void	ground, dynamic, static

Figure 10 is inference image of Ulsan Universitry with STDC2 model trained with CityScapes dataset. Inside of red lines should be marked as sidewalk or pedestrian road, but the model is marking that as road due to many road blocks are m asked roads in europe.

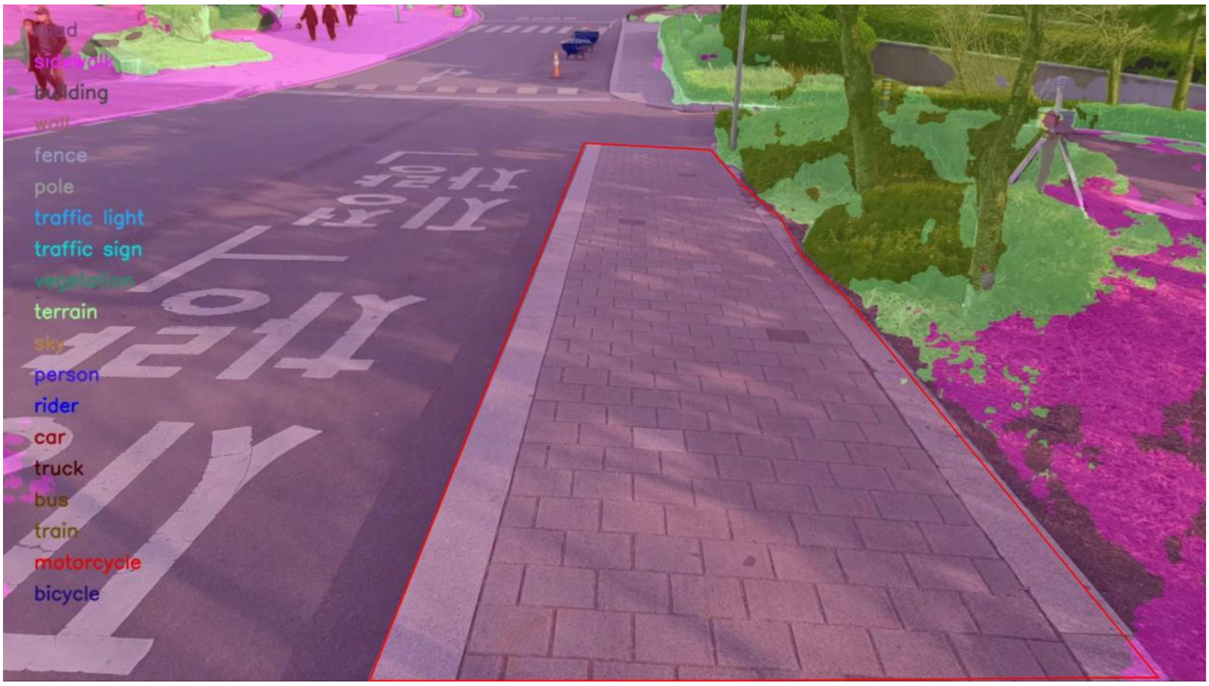


Figure 10. Inference Image with STDC2 trained with CityScapes dataset on Korean pedestrian environment

4.2 Korean Pedestrian Dataset

The image dataset is constructed of annotations for 29 types of objects that pose potential hindrances to pedestrian's sidewalk. The annotations are provided in the form of bounding boxes and polygons. Furthermore, the dataset includes information about the sidewalk surface conditions, which is also annotated using polygons.

The primary objective of developing this dataset was to enhance the mobility rights of people with disabilities. By leveraging artificial intelligence, the dataset aimed to ensure safer and smoother navigation for disabled individuals by identifying various obstacles (such as cars, people, trees, and streetlights) and damaged or hazardous walking surfaces.

Currently, there is a significant shortage of publicly available datasets tailored specifically for pedestrian AI applications, particularly in the context of Indian environments. There is a pressing need to create a domestic dataset that accurately represents the unique characteristics of Indian streets, sidewalks, and living environments. By constructing this dataset, we intend to fill this gap and provide a valuable resource for researchers and engineers working on autonomous driving applications, not only for vehicular traffic but also for diverse road types, such as residential streets and sidewalks.

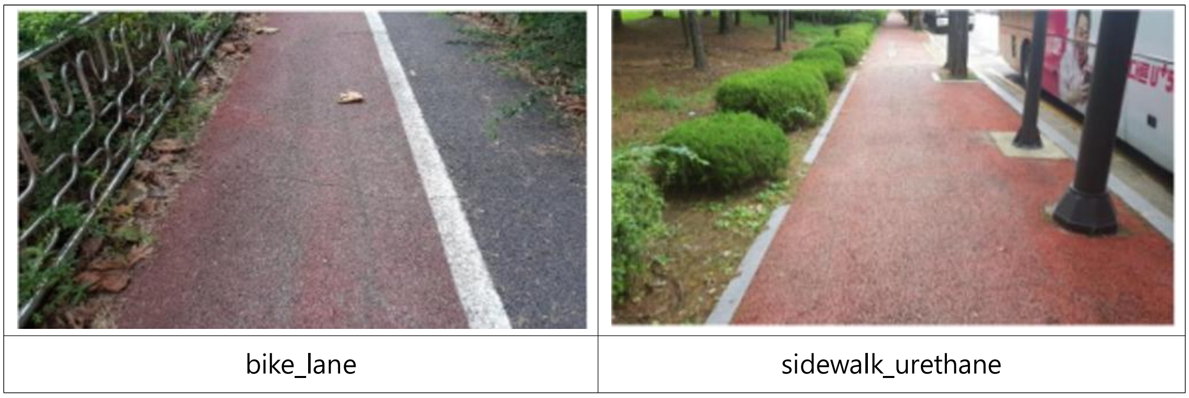
Table 2. Categories of Korean Pedestrian dataset

Category	Label
alley (road for vehicles / human)	crosswalk, damaged, normal, speed_dump
bike_lane	bike lane
braille_guide_blocks	Damaged, normal
caution_zone	Grating, manhole, repair_zone, stairs, tree_zone
roadway (only for vehicles)	crosswalk , normal
sidewalk	asphalt, blocks, cement, damaged, other, soil_stone, urethane

4.3 Training with Korean pedestrian dataset

During my initial attempt to train a semantic segmentation model, I encountered a challenge where the training did not proceed as expected; both the training and validation loss failed to converge. Upon closer examination of the dataset, I discovered substantial feature redundancy and inconsistent labeling—certain labels had visually similar features.

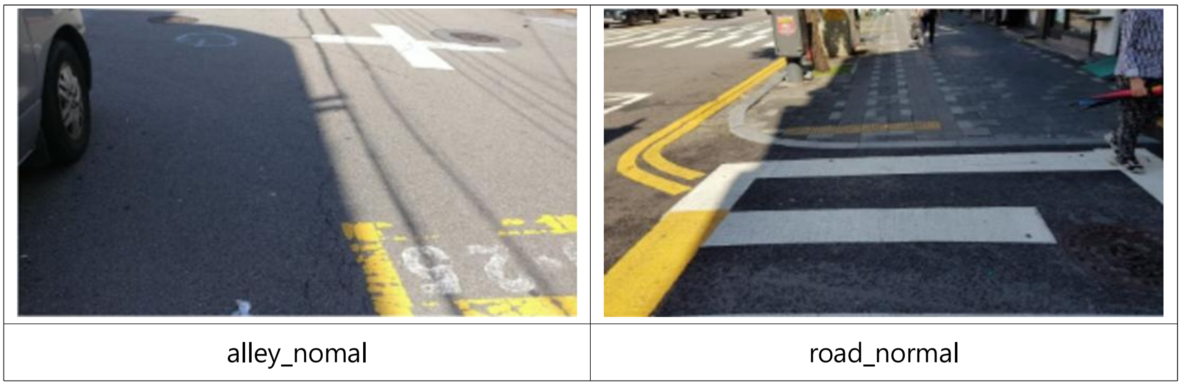
In order to address this issue, I decided to consolidate labels that shared similar visual features. As an example, all labels containing asphalt road visual features were amalgamated under the label 'alley_normal'. This new label comprised all road and alley subcategories, as well as the 'sidewalk_asphalt' subcategory, due to the identical visual characteristics shared between vehicle roads and asphalt sidewalks.



bike_lane

sidewalk_urethane

Fig 11. Example of feature redundancy in Korean Pedestrian dataset 1



alley_nomal

road_normal

Fig 12. Example of feature redundancy in Korean Pedestrian dataset 2

The label 'sidewalk_urethane' was created to include 'bike_lane' and 'sidewalk_urethane', given their feature similarities. Similarly, 'sidewalk_block' was designed to incorporate normal and damaged 'braille_guide' labels along with 'sidewalk_blocks'. Other new labels such as 'sidewalk_other', 'sidewalk_cement', 'caution_zone_tree_zone', 'caution_zone_manhole', and 'caution_zone_grating' were introduced to better categorize the dataset based on visual similarity. However, 'repair_zone' and 'stairs' were excluded due to their scarcity in the dataset.

Table 3. After label merging categories of Korean Pedestrian dataset

after merged	before merged
alley_normal	alley_rosswalk, alley_damaged, alley_normal, alley_speed_dump, road_crosswalk , road_normal, sidewalk_asphalt
sidewalk_urethane	bike_lane, sidewalk_urethane
sidewalk_blocks	braille_guide_damaged, braille_guide_normal, sidewalk_blocks,
sidewalk_ other	sidewalk_other, sidewalk_ soil_stone, sidewalk_damaged
sidewalk_ cement	Sidewalk_cement
caution_zone _tree_zone	tree_zone
caution_zone _manhole	manhole
caution_zone _grating	grating
ignored	repair_zone, stairs

Upon reclassifying the dataset, I retrained the model and used the same image for inference. However, the results remained unsatisfactory. Further analysis of the dataset revealed that the dataset's viewpoint significantly differed from the pedestrian's viewpoint. In real-life scenarios, a pedestrian's view is typically forward-facing, not looking downwards. This discrepancy led me to consider applying an image transform on the pedestrian dataset to better emulate the human perspective.

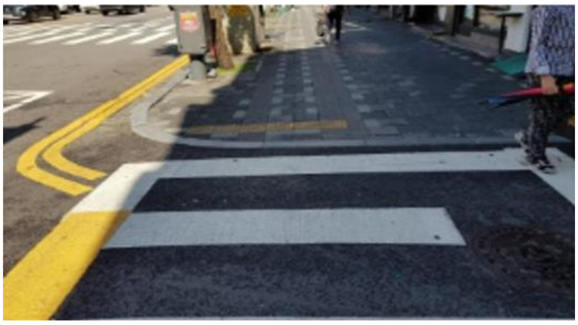
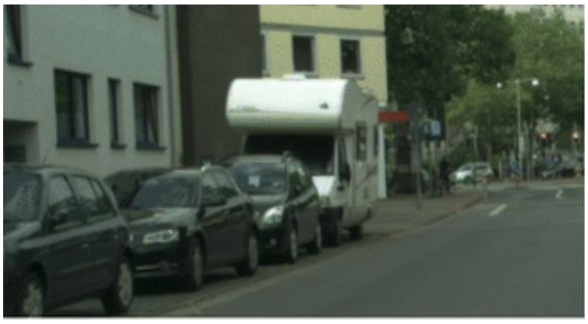
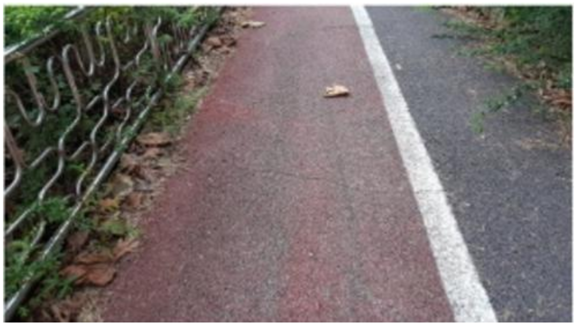
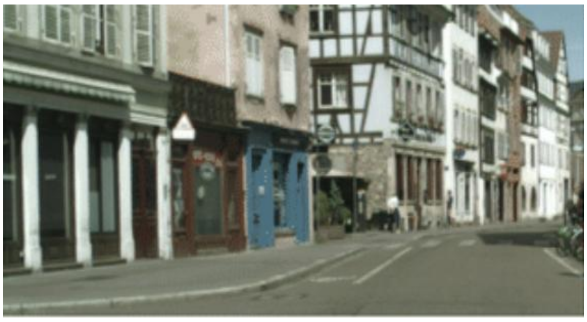
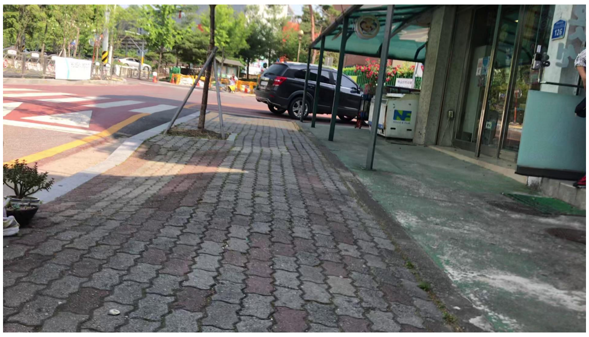
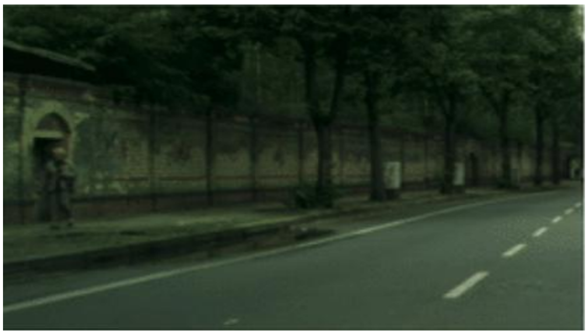
	
	
	
Korean Pedestrian	CityScapes

Table 4. Perspective comparison between datasets

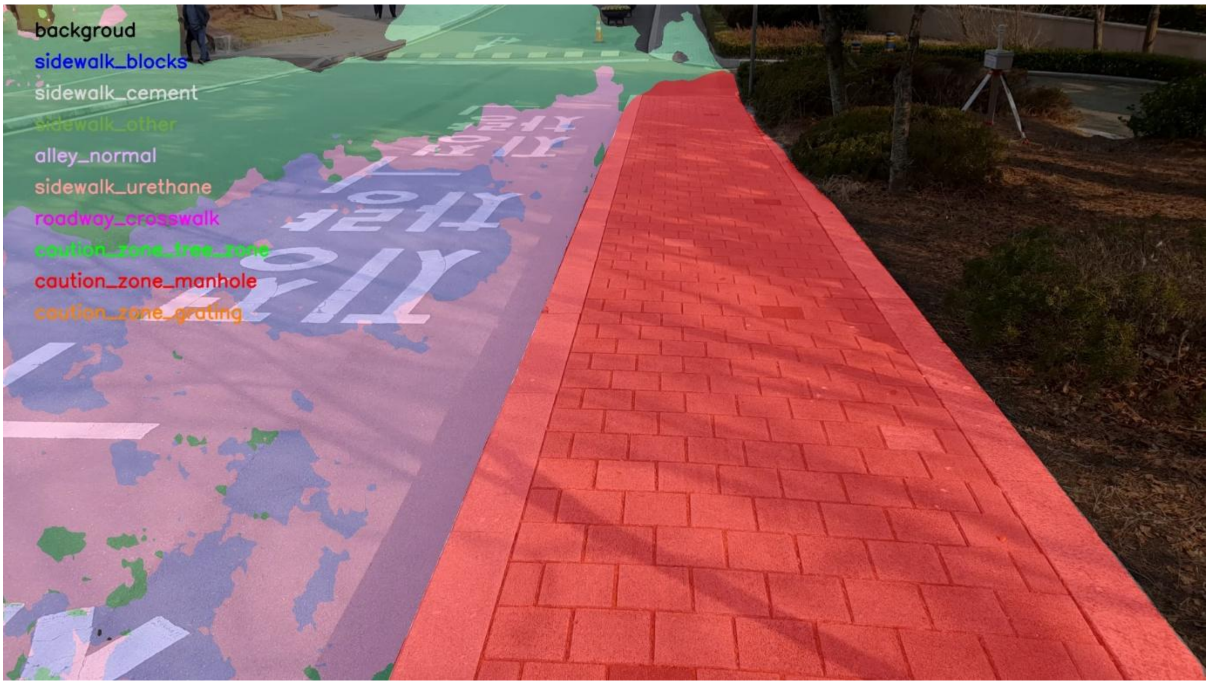


Fig 13. After label merging, the result of STDC2 trained on Korean Pedestrian dataset

4.3. Warp Perspective Transform

The concept of perspective transformation was introduced to adjust the original 1080p images. This transformation changes the x, y coordinates from 0,0 to 512, 512 and 1920, 0 to 1408, 512, yielding the transformed images as depicted in 'B'.

Upon the completion of initial training and subsequent inference with a single image using perspective transformation, the results were slightly less satisfactory compared to those obtained with the original image. Consequently, I decided to stack these transformed images to enhance the training efficiency and to provide better data regularization.

Ultimately, it was observed that the validation Mean Intersection over Union (mIoU) metrics for the transformed images outperformed those trained using the original images. This indicates that perspective transformation is a valuable technique in improving semantic segmentation performance.

Table 5. Comparison of Warp Perspective Transform

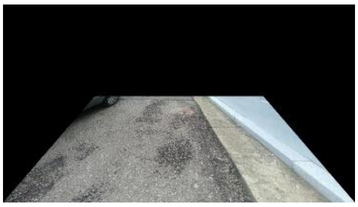
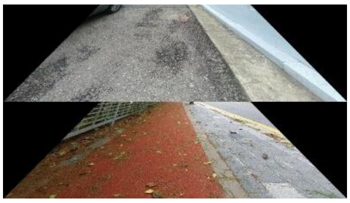
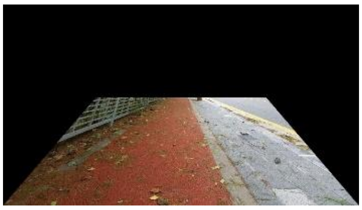
(a) Original Image	(b) Perspective Transformed	(c) b + stack
		
		

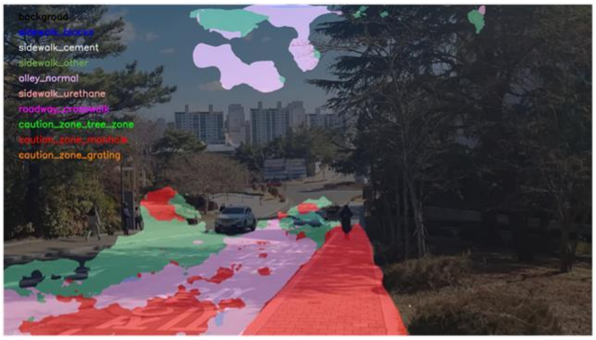
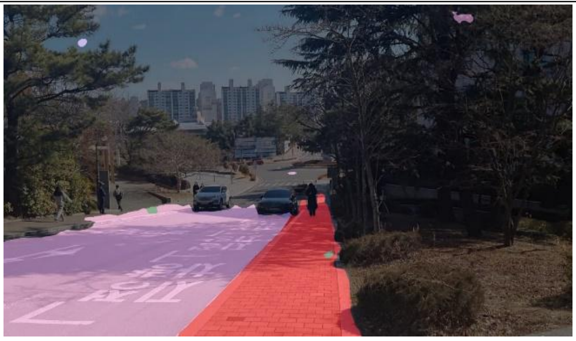
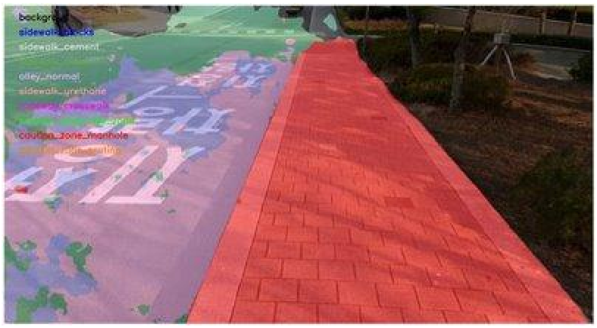
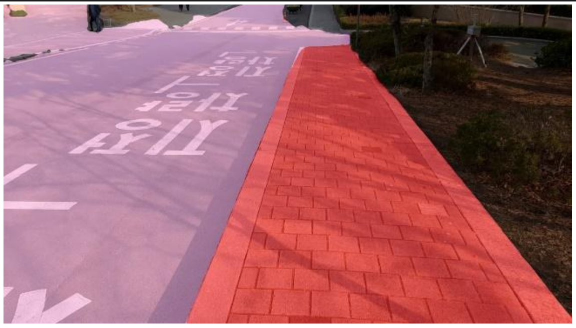
Table 6. Training result comparison with stacked / non stacked images

	a	b	c
Epoch	32,000	32,000	32,000
Added images after perspective transform	0	42,000	21,000
Number of images	42,000	84,000	63,000
mIoU	47.9%	46.24%	48.38%

Table 7. IoU of each class by stacked / non stacked images

class	a	b	c
alley_normal	62.22	63.51	67.22
sidewalk_urethane	64.79	62.49	64.79
sidewalk_blocks	64.5	62.65	67.38
sidewalk_ other	10.23	9.24	10.12
sidewalk_ cement	9.85	8.76	9.7
caution_zone _tree_zone	50.36	49.58	49.61
caution_zone _manhole	64.69	62.01	64.51
caution_zone _grating	53.76	51.76	53.71

Table 7. Result comparison of Perspective transform training

Original image training(47.9%)	Perspective transformed training(48.3%)
	
	

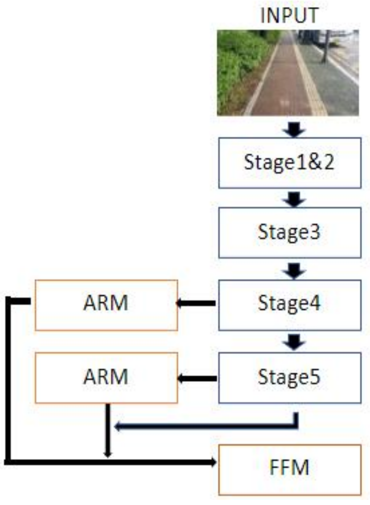
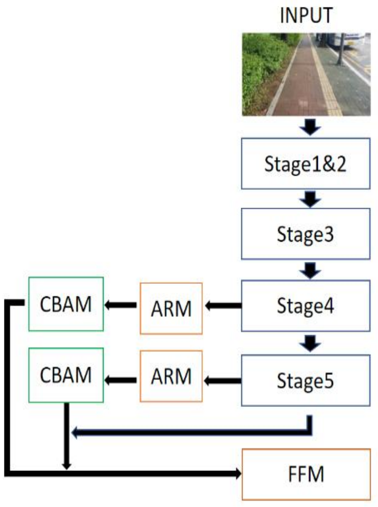
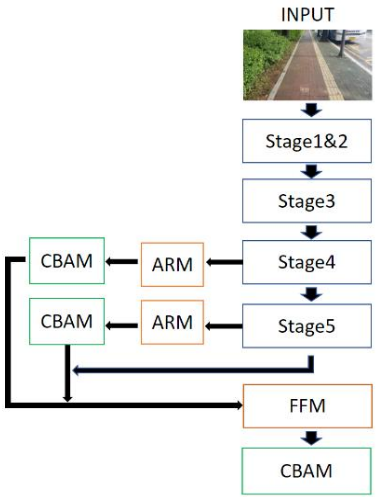
5. Experiment

5.1 Proposed Method

In this section, we will investigate the performance variations of different applications of the CBAM on the STDC2 module. As we have previously discussed, the STDC2 module integrates the ARM and the FFM, both of which focus on channel attention extraction. As such, the direct application of the SENet on this network may not result in a significant performance boost, due to the potential for redundant computations. To address this, we instead applied CBAM to the STDC2 module.

Our evaluation will comprise a comparison of two cases using the original STDC2 model: (1) integrating the CBAM module solely after the ARM module, and (2) incorporating CBAM into both the ARM and FFM modules. The schematic diagram of the modified structure is as depicted below. To facilitate a more precise examination of the model manipulations, we utilized the CityScapes dataset for comparison purposes.

Table 8. Experiment model structure

(a) Original STDC2	(b) CBAM after ARM	(C) CBAM after ARM / FFM
 INPUT Stage1&2 Stage3 ARM Stage4 ARM Stage5 FFM	 INPUT Stage1&2 Stage3 CBAM ARM Stage4 CBAM ARM Stage5 FFM	 INPUT Stage1&2 Stage3 CBAM ARM Stage4 CBAM ARM Stage5 FFM CBAM

5.2 Experiment

5.2.1 Environment and setting

Table 9. Environment and setting

Categories	Value
Operating System	Ubuntu 20.04
GPU	NVIDIA RTX 3090 X 4
CPU	Intel 10900X
Epoch	16,000
Image Size	1024 pixel

5.2.2 Result

As illustrated in Table 10, applying the CBAM after both the ARM and FFM modules resulted in a significant increase in accuracy. While this does lead to a slight decrease in inference speed, the trade-off is deemed acceptable considering the boost in accuracy. Conversely, applying the CBAM only after the ARM or FFM module did not result in a meaningful increase in accuracy. These cases only showcased an increase in their inference speed and number of computations.

Table 10. Experiment Result

	Numberof parameter	GFLOPs	mIoU(%)	Inference speed (fps)
Original STDC2	12.82M	94.26	74.03	33.87
CBAM after ARM	13.53M	95.8	73.80	33.12
CBAM after FFM	14.61M	97.31	74.57	33.08
CBAM after Both	15.32M	101.41	76.71	32.55

Table 11. Class accuracy for each experiment case (%)

class	Original	CBAM/ARM	CBAM/FFM	CBAM/Both
road	95.02	94.79	85.56	98.2
sidewalk	82.59	82.36	83.13	85.13
building	88.65	88.42	89.19	91.98
wall	50.59	50.36	51.13	54.79
fence	58.16	57.93	58.7	60.61
pole	58.33	58.1	58.87	60.78
traffic light	63.86	63.63	64.4	68.12
traffic sign	72.94	72.71	73.48	76.39
vegetation	89.66	89.43	90.2	91.87
terrain	61.26	61.03	61.8	64.48
sky	92.01	91.78	92.55	94.23
person	77.56	77.33	78.1	79.77
rider	56.55	56.32	57.09	59.47
car	94.08	93.85	94.62	94.85
truck	78.31	78.08	78.85	79.27
bus	84.41	84.18	84.95	86.53
train	72.2	71.97	72.74	73.87
motorcycle	58.23	58.0	58.77	61.22
bicycle	74.18	73.85	74.72	75.95
mIoU	74.03	73.80	74.57	76.71

6. Conclusion

In this paper, we have demonstrated how existing datasets can be effectively adapted and utilized for training real-time semantic segmentation models specifically tailored for pedestrian guidance systems. Creating a new dataset is often a time-consuming and costly process in machine learning; therefore, leveraging an existing dataset and applying perspective transformation techniques to suit the requirements of a pedestrian guidance system is a significant contribution to the field.

Furthermore, we have implemented an efficient attention mechanism that improves inference accuracy while maintaining a relatively low computational cost. This approach enables the development of a robust and accurate real-time semantic segmentation model that can be deployed on edge devices, making it more feasible and practical for real-world implementation.

The combination of these two techniques results in a powerful and efficient pedestrian guidance system, providing enhanced assistance to individuals with cognitive impairments and improving overall urban mobility. Our work paves the way for future research and development in the field of pedestrian guidance systems, with the potential to impact the lives of millions of individuals worldwide.

Reference

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. [Bibtex]

[2]https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=189

[3] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, Enhua Wu, "Squeeze-and-Excitation Networks" journal version of the CVPR 2018 paper, accepted by TPAMI, arXiv:1709.01507

[4] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon, "BAM: Convolutional Block Attention Module", ECCV 2018, arXiv:1807.06521

[5] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, Xiaolin Wei, "Rethinking BiSeNet For Real-time Semantic Segmentation", CVPR 2021, arXiv:2104.13188

[6] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, Nong Sang, "BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation ", ECCV 2018. arXiv:1808.00897

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need", NeurIPS 2017. arXiv:1706.03762

[8] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, Tat-Seng Chua, "SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning" CVF 2017, arXiv:1611.05594

[9] https://machinelearningmastery.com

[10] Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena, "Self-Attention Generative Adversarial Networks", arXiv:1805.08318

[11] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, Chang Xu, "GhostNet: More Features from Cheap Operations", CVPR 2020, arXiv:1911.11907

[13] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, Jiaya Jia, "ICNet for Real-Time Semantic Segmentation on High-Resolution Images", ECCV 2018, arXiv:1704.08545

[14] Zifeng Wu, Chunhua Shen, Anton van den Hengel, "Real-time Semantic Image Segmentation via Spatial Sparsity", CVPR, 2017, arXiv:1712.00213

[15] François Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions", CVPR 2016, arXiv:1610.02357

[16] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, Eugenio Culurciello, "ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation", CVPR 2016, arXiv:1606.02147

[17] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", MICCAI 2015, arXiv:1505.04597