

# Research on Semantic Segmentation of Fish-Eye Images for Autonomous Driving

Hongtao Huang<sup>1</sup>, Xiaofeng Tian<sup>1</sup> and Wei Tian<sup>1,\*</sup>

<sup>1</sup>School of Automotive Studies, Tongji University, Shanghai, 201804, China

**Abstract:** Fisheye cameras, valued for their wide field of view, play a crucial role in perceiving the surrounding environment of vehicles. However, there is a lack of specific research addressing the processing of significant distortion features in segmenting fish-eye images. Additionally, fish-eye images for autonomous driving face the challenge of few datasets, potentially causing over fitting and hindering the model's generalization ability.

Based on the semantic segmentation task, a method for transforming normal images into fish-eye images is proposed, which expands the fish-eye image dataset. By employing the Transformer network and the Across Feature Map Attention, the segmentation performance is further improved, achieving a 55.6% mIOU on Woodscape. Additionally, leveraging the concept of knowledge distillation, the network ensures a strong generalization based on dual-domain learning without compromising performance on Woodscape (54% mIOU).

**Keywords:** Autonomous driving, Fish-eye images, Semantic segmentation, Dual-domain learning.

## 1. INTRODUCTION

Fisheye cameras offer significant application value in autonomous driving perception tasks owing to their wide field of view. Compared with normal images, fisheye images show significant distortion, making conventional image processing methods unsuitable for fisheye images.

Traditional methods for processing fisheye images involve calibrating the fisheye image using the geometric model of the fisheye camera to correct its distortion effect, as illustrated in Figure 1(a)-(c). However, this method achieves only partial undistortion, as highlighted in Figure 1(c). Additionally, the undistorted image obtained by this method often brings new distortions, as shown in Figure 1(d).

Due to various challenges, calibrating fisheye images is insufficient to fully transform them into normal images while preserving the original field of view. As a result, numerous studies have concentrated on processing fisheye images directly, undertaking tasks such as target detection and semantic segmentation.

The transfer learning from normal to fisheye images is also of research interest. Current model training often utilizes pre-trained models for transfer learning. However, for fisheye vision tasks, the insufficient fisheye dataset size and the fact that most pre-trained



**Figure 1:** Fisheye images and their calibration [1].

models are trained on normal image datasets (e.g., the ImageNet) often hinder achieving optimal performance when transferring networks pre-trained on normal images to fisheye images.

To compensate for the scarcity of real fisheye image datasets, normal images can be transformed into fisheye images using a geometric model of fisheye

\*Address correspondence to this author at the School of Automotive Studies, Tongji University, Shanghai, 201804, China; E-mail: tian\_wei@tongji.edu.cn

camera imaging, or fisheye images can be generated on a simulator [2]. However, the generated fisheye images still differ slightly from real ones, necessitating the combination of the virtual dataset with the normal dataset for data expansion and improve model generalizability. Therefore, visual tasks for fisheye images often fall within the domain adaptation research area.

**2. RELATED WORK**

**2.1. Semantic Segmentation Algorithms for Normal and Fisheye Images**

The predominant approaches in normal image semantic segmentation algorithms are the Fully Convolutional Network (FCN) [3]. Common network structures include FCN [3], U-net [4], PSPnet [5], etc.

Additionally, the use of Transformer in semantic segmentation tasks is gaining popularity, with some works integrating the convolutional neural networks into Transformers, resulting in models such as PVT [6], Segformer [7], HRViT [8].

**Table 1: General Semantic Segmentation Networks and their Accuracy Measured in mIOU Metric**

Name	Dataset	mIOU (%)
FCN	PASCAL 2012	62.2
	Cityscapes	65.3
	ADE20K	39.3
Unet	PASCAL 2012	72.7
PSPNet	PASCAL 2012	85.5
	ADE20K	55.4
DeepLabv1	PASCAL 2012	66.4
DeepLabv3	PASCAL 2012	85.7
	Cityscapes	81.3
PVT	ADE20K	48.7
Segformer	Cityscapes	84.0
	ADE20K	51.8
HRFormer	Cityscapes	82.6
	ADE20K	57.7
	PASCAL-Context	58.5

Hanisch *et al.* [9] pioneered the fisheye image semantic segmentation research by employing the traditional segmentation algorithm (SEEDS). This approach combines feature extraction and classification, breaking down the segmentation task into three steps.

Due to the lack of specialized semantic segmentation datasets for fisheye images, early studies often transformed normal datasets into fisheye datasets using the fisheye camera model. For instance, some works [10]-[12] augmented the Cityscapes dataset through the radial geometric transformation and tested it with various network architectures including FCN, ERFnet, PSPnet, etc.

Deng *et al.* [13] utilized a deformable convolution with fixed intermediate sampling positions and adaBN to distinguish between real fisheye images and virtual fisheye images. Ye *et al.* [14] investigated the effect of datasets with different parameter transformations on model training.

However, the absence of validation using real fisheye data in most early works raises questions about the effective transfer of these methods to real fisheye images.

**2.2. Across Feature Map Attention**

The Across Feature Map Attention module (AFMA) [15] utilizes the original image, the feature map of the model's middle layer, and the final output for attention computation. It is defined as follows:

$$output = concatenate \left( P_{img} \cdot P_f^{i^T} \cdot P_{out}^i \right) + P_{out}, \tag{1}$$

where  $P_{img}$  is the segmented original image. The segmented feature map is classified by the convolution of  $N$  channels, and the downsampled feature map is  $f_a$ .  $HiWi$  is the size of the  $i^{th}$  feature map, which is then segmented to get  $P_f^i$ .  $P_{out}$  is the output of the model, and  $P_{out}^i$  is the pooling of  $P_{out}$  to  $HiWi$  size.

The Across Feature Map Attention aims to improve the results of semantic segmentation by focusing on classes similar to small objects in the image. This enhancement contributes to the model's performance in accurately segmenting small objects. Fisheye images, with their large field of view and distortion characteristics, often contain many small objects at the edges, making accurate small object recognition crucial. We propose that adding AFMA module to the network can enhance the performance in fisheye image segmentation tasks.

**2.3. Knowledge Distillation**

Knowledge distillation is proposed by Hinton *et al.* [16], and is used to compress model size, as well as

speed up training. In knowledge distillation, the hidden or output layer of the teacher model is used as a target for the student network to learn so that it can achieve similar performance as the larger model.

Knowledge distillation can be divided into two main approaches, one for the output distillation of the model and one for the intermediate layers.

The main purpose of distillation of the output of the model is to make the model directly mimic the final output of the teacher network, the distillation loss can be defined as:

$$L_{KD} = L(z_t, z_s), \quad (2)$$

where  $L$  is the KL scatter.  $z_t, z_s$  are the outputs of the teacher and the student, respectively. And the KL loss can be expressed as:

$$KL \text{ loss} = \sum P_{True}(x) \log \frac{P_{True}(x)}{P_{Pred}(x)}, \quad (3)$$

The distillation of the middle layer of the network was first proposed in Fitnets [17], suggesting an idea of matching the activation layers of the teacher-student network features. The feature-based knowledge distillation can be formulated as:

$$\begin{aligned} L_{KD}(f_t(x), f_s(x)) \\ = L(\Phi_t(f_t(x)), \Phi_s(f_s(x))), \end{aligned} \quad (4)$$

where  $f_t(x)$  and  $f_s(x)$  are the intermediate layers of the teacher and student networks,  $\Phi_t(f_t(x))$  and  $\Phi_s(f_s(x))$  refer to the transformation function in case of a mismatch in the shape of the feature maps of the teacher-student network, and  $L$  is the l2 -norm distance, or MSE loss:

$$MSE \text{ loss} = \frac{1}{N} \sum_i^N (y_i - x_i)^2. \quad (5)$$

### 3. METHODS

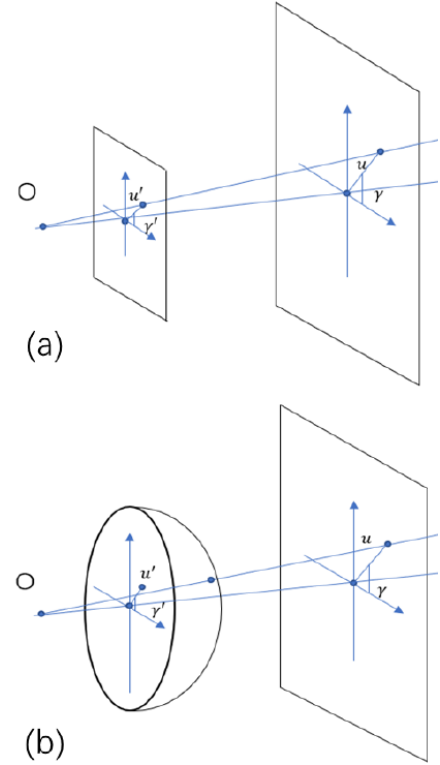
#### 3.1. Fisheye Camera Imaging Model and Transformation

The camera model defines the geometric relationship between incident light and the imaging position. In the imaging plane of a pinhole camera (Figure 2(a)), the position of a pixel is linearly related to the incident light's position in the real world, as

described by the following equations:

$$u' = ku, \gamma' = \gamma, \quad (6)$$

where  $u$  is the intersection point of the incident light in the camera plane;  $\gamma$  is the angle from the intersection point to the coordinate axis of the camera plane;  $u'$  is the pixel position of the incident light in the imaging plane;  $\gamma'$  is the angle of the pixel position on the coordinate axis of the imaging plane, and  $k$  is the scale factor.



**Figure 2:** (a) Pinhole Camera Model; (b) Fisheye Camera Model.

In the fisheye camera model (Figure 2(b)), the pixel positions on the imaging plane do not have a linear relationship with the real-world light incident position. Instead, they satisfy a geometric mapping relation  $\Phi$  described by the spherical model:

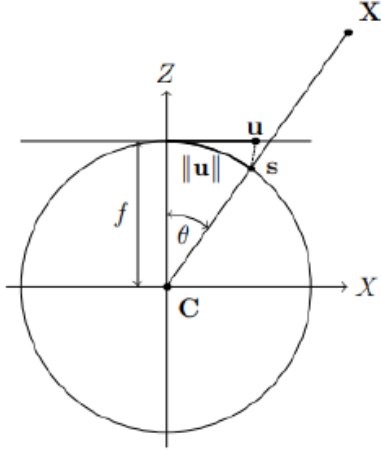
$$u' = \Phi(u), \gamma' = \gamma. \quad (7)$$

Various versions of this mapping relationship exist in related studies [18], and our work explores and experiments with the following three models:

**Equidistant model** (Figure 3). The polar distance  $u$  of the pixel position is the arc length  $\|u\|$  of the angle of the incident ray on the spherical plane:

$$\phi(u) = \|u\| = f\theta, \quad (8)$$

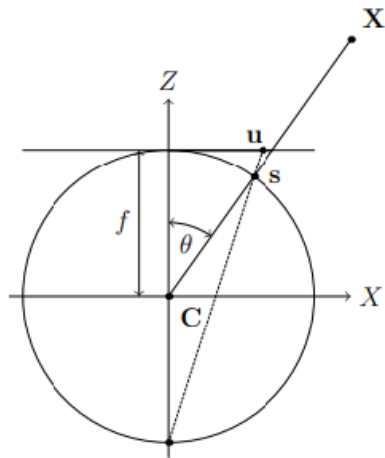
where  $u$  is the distance from the pixel to the center of the imaging plane,  $\theta$  is the angle of the incident ray to the center of light, and  $f$  is the distance from the imaging plane to the center of light.



**Figure 3:** Equidistant Model [1].

**Stereographic model** (Figure 4). The polar distance  $u$  at the center pixel position can be expressed as:

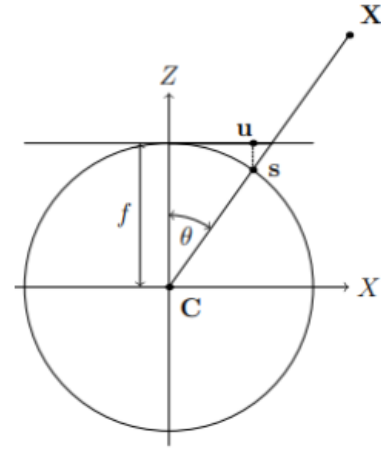
$$\phi(u) = 2f \times \tan\left(\frac{\theta}{2}\right). \quad (9)$$



**Figure 4:** Stereographic Model [1].

**Orthogonal model** (Figure 5). The polar distance  $u$  at the pixel position is the projection of the incident light at the point  $s$  on the circular arc:

$$\phi(u) = f \times \sin(\theta). \quad (10)$$



**Figure 5:** Orthogonal Model [1].

### 3.2. Fisheye Knowledge Distillation

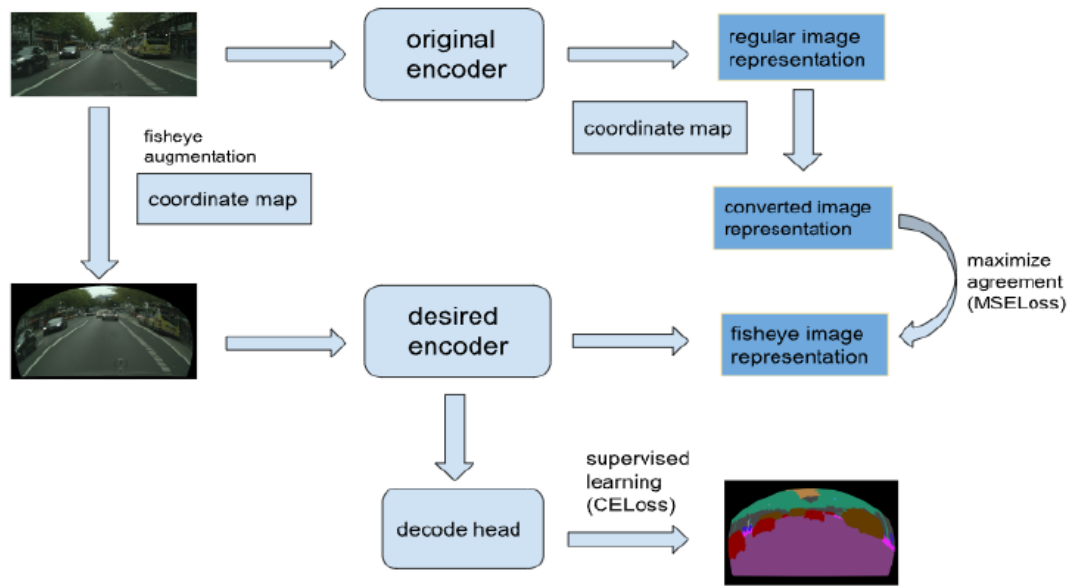
Compared to convolutional network-based models, Transformer-based models utilize the entire image for computation instead of fixed convolutional kernels. As a result, Transformer models have smaller inductive bias and larger receptive fields, making them more suitable for tasks such as semantic segmentation, which require long-distance dependencies. While many well-performing models in the Woodscape benchmark [19] utilize the attention mechanism, our focus is on exploring the performance of networks with small inductive bias properties, such as Segformer, on fisheye vision tasks.

### 3.3. Effects of Inductive Bias on Fisheye Visual Tasks

We propose a knowledge distillation approach to improve the network in learning the representation of fisheye images.

In the student-teacher model (illustrated in Figure 6), the teacher model processes a standard image, while the student model learns from the augmented fisheye image. The coordinate map of the data augmentation is passed to the network's hidden layer, where the teacher model's feature map or soft labels undergo transformation to align with the fisheye image. Subsequently, the student network learns about these features.

During the training process, the teacher network needs to be frozen. Additionally, since the coordinate mapping map can only be transformed for images of the same size, the hidden layer of the teacher network is upsampled to match the size.



**Figure 6:** Fisheye Distillation Learning Process.

### 3.4. Dual-Domain Learning

As mentioned in Section 1, the domain adaptation combining virtual and real data is a major focus of semantic segmentation research. Our focus is on domain adaptation learning using the Cityscapes virtual fisheye dataset and the Woodscape real fisheye dataset, *i.e.*, by combining the two datasets for training.

Ideally, larger datasets pose a greater challenge for model fitting, leading to higher model metrics. However, the combination of the two datasets, virtual and real, does not necessarily improve metrics due to the presence of a domain gap. For instance, in Synwoodscape 6, the authors trained the virtual images generated on the simulator in combination with real images but failed to achieve higher metrics.

In our work, we engage in dual-domain learning on virtual and real data. However, the categories labeled by Cityscapes and Woodscape in differ from each other. This necessitates the network to learn different categories in the two domains. Therefore, our approach involves dual-classification head learning, with each classification head trained on its respective dataset. Finally, we fine-tune the network.

## 4. EXPERIMENTS

### 4.1. Dataset

This section outlines the datasets utilized in the experiments of our work.

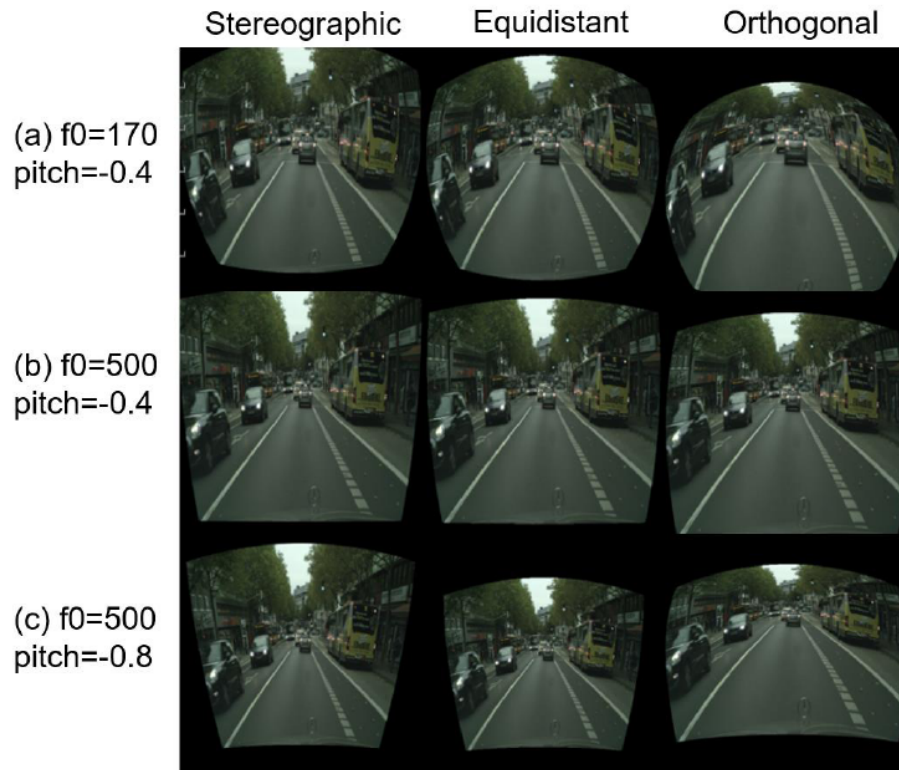
Woodscape [20] is a real fisheye dataset provided by ValeoAI in 2021. It comprises 10,000 labeled images for semantic segmentation, with 8,200 images constituting the public dataset and the remaining 1,800 forming the non-public test set. The dataset encompasses tasks such as semantic segmentation, instance segmentation, 2D target detection, mud dirt detection, and end-to-end driving data. The semantic segmentation annotations encompass nine categories (excluding the null category) and include four views of the car from the front, back, left, and right.

Cityscapes is a standard semantic segmentation dataset for driving scenes, consisting of driving recorder images from 50 different cities. It includes 5,000 finely labeled images and 20,000 roughly labeled images, covering a total of 19 categories.

### 4.2. Transforming Images from Pinhole to Fisheye Camera

Comparing with Figure 7(a)(b), changing the focal length in the orthogonal model results in the most noticeable radial distortion. Similarly, comparing with Figure 7(b)(c), keeping the focal length unchanged in the orthogonal model reveals the most significant distortion when altering the viewing angle pitch of the imaging model. In order to mimic the large distortion characteristics in fisheye images, all subsequent experiments in this work use the orthogonal model as a benchmark.





**Figure 7:** Performance of different fisheye camera models on pinhole camera image transformation.

**4.3. Performance Comparison of Different Network Transfer**

To compare the performance of different pre-trained models on fisheye images and determine the baseline model for our work, various typical network structures are selected. Pre-training weights are obtained from publicly available sources on Cityscapes, and the

models are tested on Woodscape. Subsequently, their metrics are compared on both the Cityscapes dataset and Woodscape.

Comparing various pre-trained models with similar metrics on the Cityscapes dataset (Table 2), Segformer [7], featuring a Transformer structure, exhibits superior performance on Woodscape. This could be attributed

**Table 2: Comparison and Visualization of Transfer Performance (Measured by mIOU in %) for Different Structured Networks**

Net \ Dataset	Deeplabv3	Erfnet	HRnet	Segformer
Cityscapes	76.7	74.75	81.6	76.54
Woodscape	22.3	20.8	25.12	30.9
Visualization				

to improved generalization performance by the Transformer structure after extensive training, owing to the absence of the inductive bias property in convolutional neural networks. Consequently, the Segformer network excels in transfer performance compared to the other networks considered.

**4.4. Validation of Fisheye Data Augmentation**

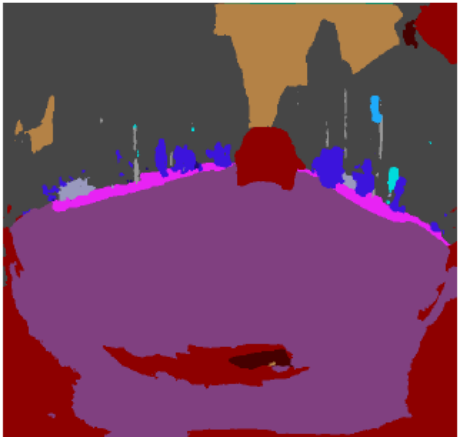
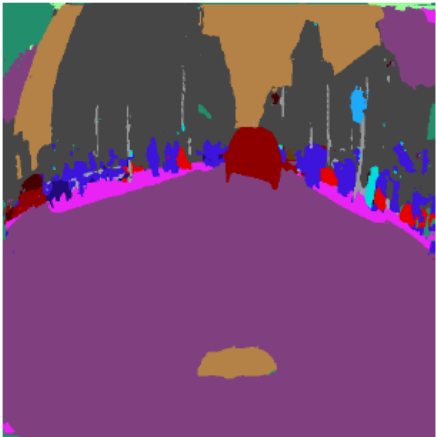
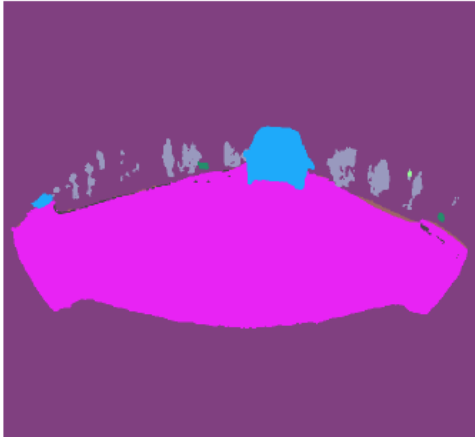
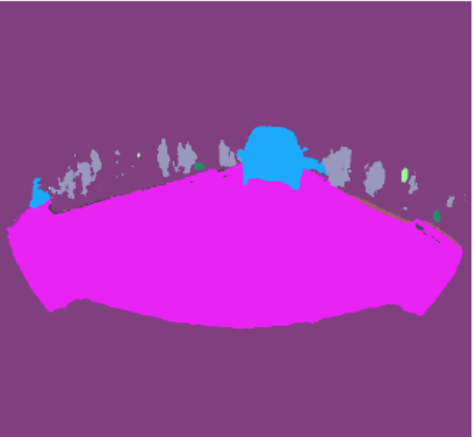
This experiment verifies the effectiveness of fisheye data augmentation.

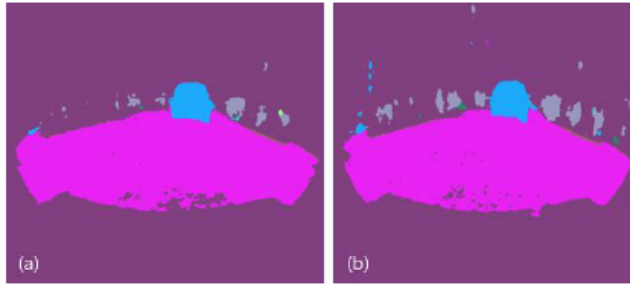
In our work, the orthogonal model is employed for online data augmentation on the Cityscapes dataset, utilizing transform parameters ranging from (f0:500~1000; pitch:0~-0.60). The augmented data is then input into the network for training. Training is

conducted with a constant learning rate, 200 warm-up steps, a learning rate set to 2e-5, an Adam optimizer, a total duration of 20 epochs, and a batch size of 32. After the training, the results of the original network and the pre-trained network are compared with those of the real fisheye dataset, and the results are fine-tuned on the real fisheye dataset.

As observed in Table 3, the pre-trained model exhibits superior performance on the real fisheye dataset, with higher performance achieved after fine-tuning. In addition, considering that the labeled classes in Cityscapes and Woodscape are not the same, it is necessary to redefine the classification head. Thus, the network's backbone is frozen, and only the decoder is trained. Subsequently, it is fine-tuned on Woodscape.

**Table 3: Comparison of Fisheye Augmentation by Fine-Tuning**

	Original Model	Virtual Fisheye Images with Pre-Trained Model
Learning on Cityscape		
Performance	Mean_IOU: 0.56	Mean_IOU: 0.572
Fine-Tuned on Woodscape		



**Figure 8:** Visualization of the model output trained (a) without fish-eye data augmentation, (b) with fish-eye data augmentation.

**Table 4: Performance of Fisheye Data Augmentation (Entire Network)**

	Mean_IOU	Mean_acc
No augmentation	0.5661	0.5804
Augmentation	0.6626	0.6639

Based on the observations from Table 4 and Figure 8, the pre-trained network with the virtual fisheye dataset demonstrates superior transfer performance on the real fisheye dataset compared to the pre-trained network with normal images.

#### 4.5. Validating the Validity of AFMA Module

**Table 5: Performance of Across Feature Map Attention Modules (Measure by mIOU in %)**

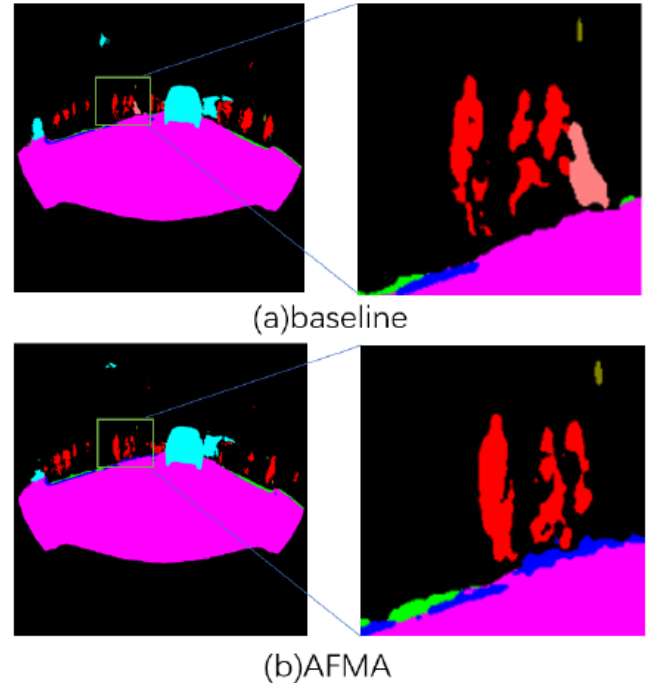
	Baseline	With AFMA
Road	0.9401	<b>0.9412</b>
lanemark	0.7020	<b>0.7068</b>
Curb	0.5543	<b>0.5547</b>
Person	0.4363	<b>0.4465</b>
Rider	0.3391	<b>0.3396</b>
Vehicles	0.8684	0.8683
Bicycle	0.4160	<b>0.4220</b>
Motorcycle	0.3745	<b>0.3798</b>
Traffic sign	0.2826	0.2662
Overall	0.5645	<b>0.5647</b>

To verify the effectiveness of small object segmentation on fisheye images, we tested the performance of the network before and after loading the AFMA module. After initial fine-tuning on Woodscape, online image augmentation is employed to prevent overfitting. This involves a random cropping with a rate of 0.8 to 1 and a random flipping. The

training utilizes the Adam optimizer with 200 warm-up steps and a learning rate of  $2e-4$ . During metrics evaluation, specific calculations are performed for each class to assess the segmentation performance of small object classes, such as lanemark, curb, and person.

As seen in Table 5, the segmentation performance of small objects such as lanemark, curb, and person are improved, and the overall metrics are also slightly improved.

A decrease in metrics for traffic signs is observed, possibly because traffic signs appear sparsely and infrequently within the same map. The module relies on pixels of the same class at other locations in the map for small object supervision, yet the scarcity of traffic signs diminishes the supervisory performance of this module for this class.



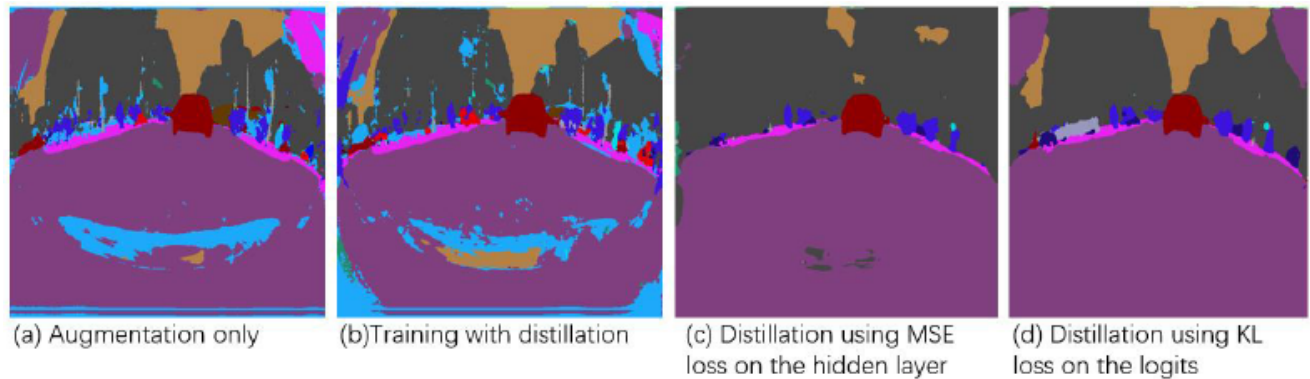
**Figure 9:** Visualization of performance by baseline and AFMA module.

#### 4.6. Exploration of Fisheye Knowledge Distillation

This experiment verifies the effectiveness of knowledge distillation. Three different training methods are mainly compared: training with data augmentation, training with distillation, and training by combining both.

In our work, we explored two distillation methods: distillation of the hidden layer using MSE loss and distillation of the soft labels directly on the network output using KL loss.





**Figure 10:** Performance of fisheye knowledge distillation learning.

Figure 10 displays the results. When using MSE loss for distillation against the hidden layer, the network is challenging to fit. On the other hand, when distilling using transformed soft labels with KL loss, the network is easier to fit but does not achieve the desired performance. The accuracy of some classes decreases in comparison to the results obtained from direct training with hard labels.

The fisheye distillation network exhibits reduced noise compared to the network without distillation. However, it does not achieve high accuracy and is harder to train.

A hypothesis is that the hidden layers or soft labels used for distillation originate from the same network, thus the output of each pixel corresponding to the same image is not transformed. The fisheye geometric transformation only changes its position, and for an image with a transformed distortion parameter, learning a fixed value for each pixel may cause the network to lose translational isotropy, thereby hindering the optimal parameter learning.

#### 4.7. Dual-Domain Learning

The experiment combines the virtually generated fisheye dataset and the real dataset to train the network on both domains, addressing the issue of overfitting observed in small fisheye image datasets.

A potential problem with training the network on both the Cityscapes virtual fisheye dataset and the Woodscape real fisheye dataset is that the labels of the two datasets are inconsistent. When training with a single classification head, it is necessary to combine the classes of the two datasets and to ignore the classes that do not exist in the other dataset; At the

same time, two classification heads corresponding to the two datasets are also used for training, and then the weights of the two classification heads are combined after training.

As can be seen in Figure 11, due to the different classes in the two datasets, the use of a single classification head needs to mask its class on the other dataset, and the method of identification based on confidence adopted in the experiments tends to leave a large number of pixels unsupervised, and therefore inaccurate segmentation can occur;

When training with two classification heads on each of the two datasets, since no masking is required, every pixel can be supervised, which is better than using a single classification head, where the classification head trained on Woodscape can achieve similar metrics (mIOU=0.56) as the baseline network (fine-tune on Woodscape only), while the merged classification head on Woodscape also reveals classes on another dataset, such as sidewalks, buildings and sky.

Based on the comparison in Figure 11, utilizing two classification heads on two datasets emerges as a more reasonable training method. This approach of dual-domain learning not only expands the dataset for the network to learn but also prevents overfitting on a single dataset. It enhances the generalization of the model, enabling the network to learn more features of the classes without compromising overall performance.

## 5. CONCLUSION

Our work centers around a series of hypothesized ideas of fisheye image segmentation tasks.

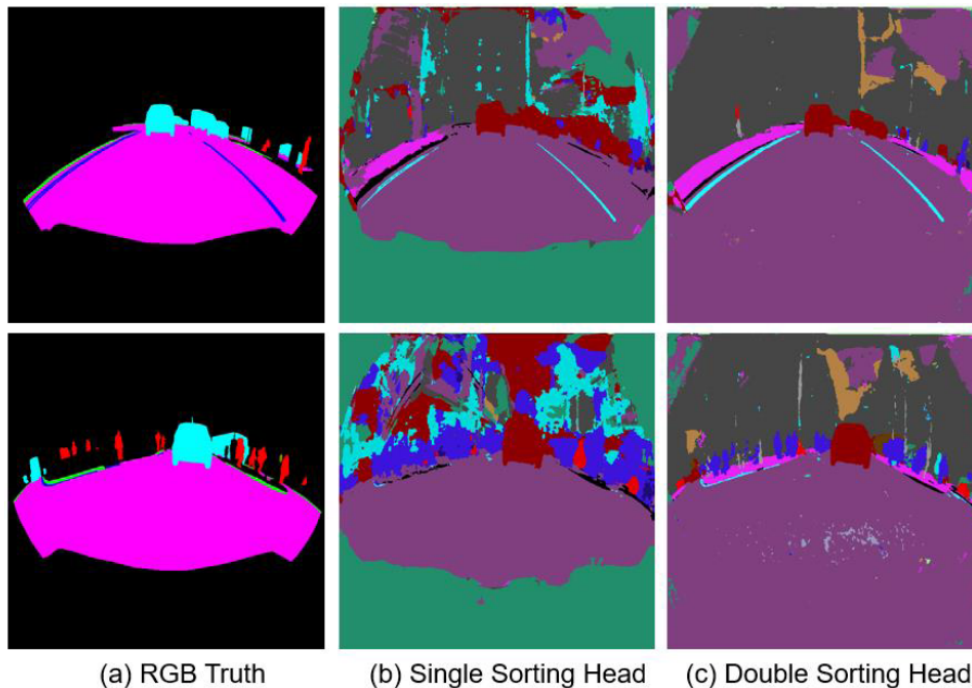


Figure 11: Dual-domain learning visualization.

Experimental results reveal that the Transformer network structure exhibits superior robustness and transfer performance on fisheye images compared to convolutional networks. Data augmentation using a fisheye camera model and pre-training on the generated virtual fisheye dataset improves the model's effectiveness on real datasets and makes fine-tuning easier to fit.

Meanwhile, the Across Feature Map Attention module, designed to aid the segmentation of small objects, proves effective in enhancing the network's performance on fisheye images. Fisheye distillation learning reduces image noise, but the results, while less noisy, are not as accurate as those obtained through direct training. Finally, dual-domain learning with two classification heads enables the network to learn more features while maintaining metrics similar to direct fine-tuning.

REFERENCE

[1] Kumar, Varun Ravi, et al. "Surround-View Fisheye Camera Perception for Automated Driving: Overview, Survey & Challenges." IEEE Transactions on Intelligent Transportation Systems 24 (2022): 3638-3659. <https://doi.org/10.1109/TITS.2023.3235057>

[2] Ekkat, Ahmed Rida et al. "SynWoodScape: Synthetic Surround-View Fisheye Camera Dataset for Autonomous Driving." IEEE Robotics and Automation Letters 7 (2022): 8502-8509. <https://doi.org/10.48550/arXiv.2203.05056>

[3] Shelhamer, Evan et al. "Fully convolutional networks for semantic segmentation." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014): 3431-3440. <https://doi.org/10.48550/arXiv.1411.4038>

[4] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015. <https://doi.org/10.48550/arXiv.1505.04597>

[5] Zhao, Hengshuang et al. "Pyramid Scene Parsing Network." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 6230-6239. <https://doi.org/10.48550/arXiv.1612.01105>

[6] Wang, Wenhai et al. "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions." 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021): 548-558. <https://doi.org/10.48550/arXiv.2102.12122>

[7] Xie, Enze et al. "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers." Neural Information Processing Systems (2021). <https://doi.org/10.48550/arXiv.2105.15203>

[8] Gu et al. "Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation." 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021): 12084-12093. <https://doi.org/10.48550/arXiv.2111.01236>

[9] Hänisch, Evangelio, Tadjine and Pätzold, "Free-space detection with fish-eye cameras," 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 135-140, <http://doi.org/10.1109/IVS.2017.7995710>

[10] Deng, Yang, Qian et al. "CNN based semantic segmentation for urban traffic scenes using fisheye camera," 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 231-236, <http://doi.org/10.1109/IVS.2017.7995725>

- [11] Sáez, Bergasa, Romeral *et al.* "CNN-based Fisheye Image Real-Time Semantic Segmentation," 2018 IEEE Intelligent Vehicles Symposium (IV), 2018, pp. 1039-1044, <http://doi.org/10.1109/IVS.2018.8500456>
- [12] Blott, G., Takami, M., & Heipke, C. (2018). Semantic Segmentation of Fisheye Images. ECCV Workshops. [https://doi.org/10.1007/978-3-030-11009-3\\_10](https://doi.org/10.1007/978-3-030-11009-3_10)
- [13] Deng, Liuyuan *et al.* "Restricted Deformable Convolution-Based Road Scene Semantic Segmentation Using Surround View Cameras." IEEE Transactions on Intelligent Transportation Systems 21 (2018): 4350-4362. <http://doi.org/10.1109/TITS.2019.2939832>
- [14] Ye, Yaozu *et al.* "Universal Semantic Segmentation for Fisheye Urban Driving Images." 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (2020): 648-655. <https://doi.org/10.1109/SMC42975.2020.9283099>
- [15] Sang, Zhou, Islam and Xing, "Small-Object Sensitive Segmentation Using Across Feature Map Attention," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 5, pp. 6289-6306, 1 May 2023, <https://doi.org/10.1109/TPAMI.2022.3211171>
- [16] Hinton, Vinyals & Dean (2015). Distilling the Knowledge in a Neural Network. ArXiv, abs/1503.02531. <https://doi.org/10.48550/arXiv.1503.02531>
- [17] Romero, Ballas, Kahou, Chassang, Gatta, & Bengio. (2014). FitNets: Hints for Thin Deep Nets. CoRR, abs/1412.6550. <https://doi.org/10.48550/arXiv.1412.6550>
- [18] Miyamoto, "Fish eye lens," Journal of the Optical Society of America, vol. 54, no. 8, pp. 1060-1061, 1964.
- [19] Ramachandran, Saravanabalagi *et al.* "Woodscape Fisheye Semantic Segmentation for Autonomous Driving - CVPR 2021 OmniCV Workshop Challenge." ArXiv abs/2107.08246 (2021) <https://doi.org/10.48550/arXiv.2107.08246>
- [20] Yogamani, Senthil Kumar *et al.* "Woodscape: A Multi-Task, Multi-Camera Fisheye Dataset for Autonomous Driving." 2019 IEEE/CVF International Conference on Computer Vision (ICCV): 9307-9317. <https://doi.org/10.1109/ICCV.2019.00940>

Received on 19-11-2023

Accepted on 14-12-2023

Published on 27-12-2023

DOI: <https://doi.org/10.31875/2409-9694.2023.10.13>© 2023 Huang *et al.*

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.