# Atmospheric Transmission and Thermal Inertia Induced Blind Road Segmentation with a Large-Scale Dataset TBRSD

Junzhang Chen[1], Xiangzhi Bai[1,2,3*]
[1]Image Processing Center, Beihang University, Beijing, China
[2]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[3]Advanced Innovation Center for Biomedical Engineering, Beihang University
chenjz@buaa.edu.cn, jackybxz@buaa.edu.cn

## Abstract

*Computer vision-based walking assistants are prominent tools for aiding visually impaired people in navigation. Blind road segmentation is a key element in these walking assistant systems. However, most walking assistant systems rely on visual light images, which is dangerous in weak illumination environments such as darkness or fog. To address this issue and enhance the safety of vision-based walking assistant systems, we developed a thermal infrared blind road segmentation neural network (TINN). In contrast to conventional segmentation techniques that primarily concentrate on enhancing feature extraction and perception, our approach is geared towards preserving the inherent radiation characteristics within the thermal imaging process. Initially, we modelled two critical factors in thermal infrared imaging - thermal light atmospheric transmission and thermal inertia effect. Subsequently, we use an encoder-decoder architecture to fuse the feathers extracted by the two modules. Additionally, to train the network and evaluate the effectiveness of the proposed method, we constructed a large-scale thermal infrared blind road segmentation dataset named TBRSD consists 5180 pixel-level manual annotations. The experimental results demonstrate that our method outperforms existing techniques and achieves state-of-the-art performance in thermal blind road segmentation, as validated on benchmark thermal infrared semantic segmentation datasets such as MFNet and SODA. The dataset and our code are both publicly available in* `https://github.com/chenjzBUAA/TBRSD` *or* `http://xzbai.buaa.edu.cn/datasets.html`.

## 1. Introduction

As of 2020, the estimated number of blind individuals worldwide was 43.3 million, and 295 million people had moderate to severe vision impairment. Between 1990 and 2020, the global number of people who were blind increased by 50.6%, while the number of individuals with moderate to severe vision impairment rose by 91.7%. Projections indicate that by 2050, the number of blind individuals will increase to 51 million, and the number of people with moderate to severe vision impairment will rise to 474 million [3]. Individuals with visual impairments often experience difficulty with navigation when engaging in social

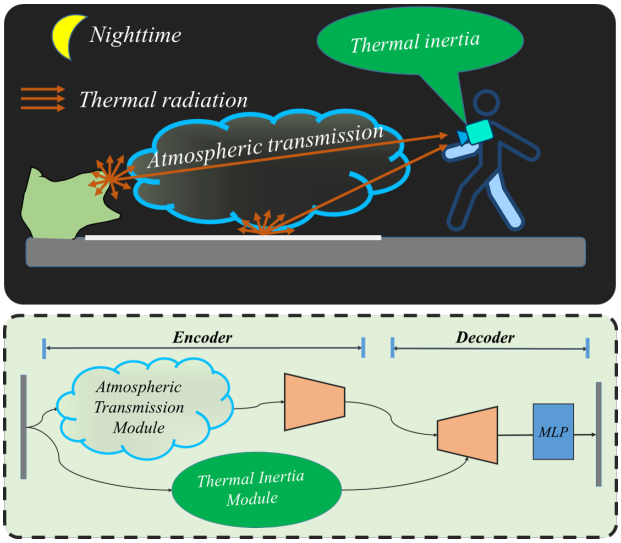


Figure 1. Thermal infrared imaging is influenced by two key factors: atmospheric transmission and thermal inertia effect. All objects with a temperature above absolute zero emit thermal infrared radiation, and the intensity of this radiation that is detected by the camera is affected by atmospheric transmission. Once the incident radiation reaches the thermal microbolometer, the resulting intensity measurement is also affected by the thermal inertia effect. Thus, we modeled the two effects and incorporated these models into our blind road segmentation method, which allows us to more accurately identify and segment thermal images of blind roads, even in challenging environmental conditions.

*Corresponding author.

activities and interactions due to the complex outdoor environment. To help tackle this problem, a variety of supportive tools and assistants have been developed to promote outdoor navigation and enhance the use of technology for those with visual impairments [11].

Since the 1960s, walking assistants have been introduced to help people with mobility issues with tasks such as navigation and location. These assistants help individuals with visual impairments detect and locate obstacles around them through the use of sensors that perceive the surrounding environment [4, 17, 39].

Visual assistance technology has spurred the development of various walking assistants to help mitigate the mobility obstacles of visually impaired individuals [14]. Various methods have been proposed to assist visually impaired individuals to walk on blind roads, such as visual guiding using image segmentation [1, 13, 12, 29, 26, 5, 19, 37].

Semantic segmentation methods can be utilized to segment blind road. Convolutional neural networks (CNNs) have shown great performance improvements in semantic segmentation [2, 6, 18, 28, 38]. Vision Transformers (ViTs) have recently emerged as a competitive alternative. ViTs offer strong expressivity and long-distance information interaction with dynamic feature aggregation through self-attention operations [40, 34, 8, 7, 16]. Furthermore, datasets containing visual images with sidewalks have been introduced to offer assistance to individuals with visual impairments, exemplified by projects like SideGuide and Mapillary Vistas [24, 22].

However, most semantic segmentation methods rely on visual images, which can be inadequate in weak illumination environments like darkness or fog. Therefore, thermal infrared semantic segmentation has been proposed as a solution to this problem [15, 31, 21, 23, 35].

In spite of the extensive utilization of thermal infrared images across various applications, the influence stemming from the thermal imaging characteristics of objects and backgrounds has frequently been disregarded by prevailing processing techniques. The prevalent practice of interpreting thermal infrared images solely as grayscale images fails to comprehensively capture the intricate thermal nuances.

While conventional segmentation methodologies predominantly center around refining feature extraction and perception, the precision of these techniques is frequently compromised due to the introduction of information disorder through the thermal imaging process. To address this limitation, we identify two pivotal factors significantly impacting thermal infrared imaging - atmospheric transmission and the thermal inertia effect - and incorporate them into the thermal infrared image processing pipeline.

Our specific proposition involves a network that capitalizes on these effects to enhance the efficacy of blind road segmentation in thermal infrared imaging. As a result, our method outperforms in the domain of thermal infrared blind road segmentation, rectifying the limitations previously encountered.

To summary, our main contributions are as follows:

- We model the two critical factors in thermal infrared imaging - atmospheric transmission and thermal inertia effect - and incorporate them into the segmentation task for the first time. we present a thermal blind road segmentation network that effectively leverages these effects.
- We construct the first public large-scale thermal infrared blind road segmentation dataset, which contains 5180 images. All images are with pixel-level annotations. The dataset is publicly available.
- Our method achieves state-of-the-art performance on both our own thermal infrared blind road segmentation dataset TBRSD and the benchmark thermal infrared semantic segmentation datasets MFNet [10] and SODA [15].

## 2. Related Work

### 2.1. Walking Assistant Systems

Assisting visually impaired individuals to navigate safely on foot is a challenging task. While traditional tools such as guide dogs [32] and white canes [36] have been useful, their effectiveness is limited by factors such as speed, coverage, and capacity. In recent years, visual assistance methods have been proposed to address these limitations, with the goal of improving the mobility and independence of visually impaired individuals. One such method is visual guiding, which uses image segmentation techniques to assist individuals in walking on blind roads. In particular, researchers such as Alvarez et al. have used an illumination-invariant feature space and road class-likelihood to construct a classifier for road detection [1], while Horne et al. have designed a semantic labeling method to assist with path navigation [13, 12]. Tang et al. proposed a blind roads segmentation method using Gaussian vectors and multi-color spaces, but it has shown poor performance in complex environments [29]. To address these limitations, Peng et al. designed a blind roads segmentation method based on a color histogram and gray level co-occurrence matrix, which has better illumination shielding effects but requires longer computation times [26]. Cao et al. proposed a lightweight semantic segmentation network to quickly and accurately segment blind roads and crosswalks. The method use depthwise separable convolution to improve the speed of network segmentation, and use a dense atrous spatial pyramid pooling module and context feature fusion module to ensure segmentation accuracy [5].

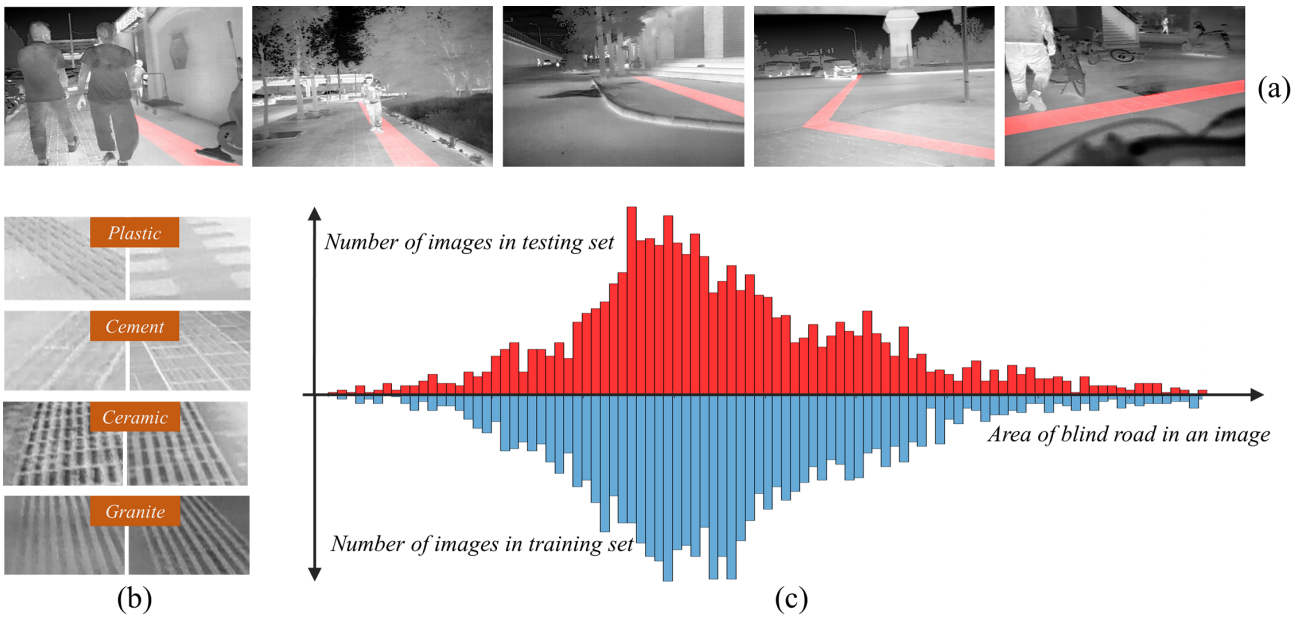


Figure 2. We construct the first public large-scale thermal blind road segmentation dataset TBRSD. Here are (a) some samples of original images and pixel-level annotations from TBRSD and (b) types of blind road materials present in the dataset. (c) The distribution of the training and testing sets in TBRSD, represented by a histogram. The horizontal axis indicates the area of blind road in an image, and the vertical axis indicates the number of images with that area.

### 2.2. Thermal Image Segmentation

Thermal image segmentation is a technique that involves learning thermal infrared image features in dedicated modules to obtain accurate segmentation results. Edge information is used to handle ambiguous object boundaries and imaging noise, which improve the performance of segmentation quality [15] . [31] proposed a thermal infrared pedestrian segmentation algorithm based on a conditional generative adversarial network (IPS-cGAN). The IPS-cGAN generator is based on the Unet network, but with two modifications to make it better suited for thermal infrared pedestrian segmentation. In RT-SegRBM-Net [21], a novel approach is proposed for real-time segmentation of vehicles in UAV-based thermal images. This approach combines the Gaussian-Bernoulli Restricted Boltzmann Machine (GB-RBM) and convolutional neural network to achieve accurate and efficient segmentation results. Feature Transverse Network (FTNet) [23] is a convolutional neural network architecture that can be trained end-to-end. Employing an encoder-decoder structure along with an edge guidance component, FTNet is designed to perform accurate pixel-wise classification. To address the challenge of thermal image segmentation in nighttime driving scenes, Xiong et al. proposed the Multi-level correction network (MCNet) [35]. This approach leverages a multi-level attention module (MAM) to effectively capture the contextual information in thermal images, enabling more accurate segmentation results.

## 3. Thermal Blind Road Segmentation Dataset

We collect a large-scale thermal blind road segmentation dataset named TBRSD. TBRSD consists of 5180 frames with manually pixel-level segmentation annotation.

The Thermal camera used in our dataset is the Zenmuse XT, equipped with an uncooled vanadium oxide microbolometer for thermal imaging. The camera has a spectral range of 7.5-13.5 $\mu m$, a pixel spacing of 17$\mu m$, and a focal length of 13$mm$.

Over 450,000 frames were meticulously captured by the camera across various geographical points in northern, central, and eastern China, each representing distinct urban architectures and diverse climatic conditions. The data collection endeavors were undertaken post 8 pm, specifically during periods of low illumination. The dataset encompasses frames acquired under a gamut of weather scenarios, ranging from sunny and rainy to cloudy and foggy, covering all seasons. All frames maintain a consistent resolution of 480$\times$720 pixels.

These images were captured while walking and handholding the camera, effectively emulating the sensory experience of individuals with visual impairments. To ensure accuracy, every pixel-level annotation underwent a comprehensive manual process and subsequent verification, employing visible images taken at the same locations during daytime. Some samples and pixel-level annotations are giv-

en in Fig. 2 (a).

Our dataset, TBRSD, contains samples of major types of materials used in blind road construction, including plastic, cement, ceramic, and granite, as shown in Fig. 2(b). The dataset is divided into train, test, and validation sets in a standard manner, with 2500, 500, and 2180 images, respectively. The distribution of the areas of blind road in the train and test sets is shown in the histogram in Fig. 2(c). The distribution of blind road areas in the images follows a normal distribution and is similar in both the train and test sets.

To the best of our knowledge, TBRSD is the first public thermal pixel-level dataset for blind road segmentation task.

## 4. Thermal Imaging Induced Blind Road Segmentation Neural Network

In this work, we propose a novel deep neural network (TINN) that takes into account two key effects in thermal imaging, namely the atmospheric transmission of thermal radiation and the thermal inertia effect. Instead of regarding thermal images solely as grayscale depictions and focusing on enhancing feature extraction and perception capabilities, we have devised an encompassing model that effectively encapsulates the two crucial physical phenomena that wield substantial influence on thermal imaging.

To this end, we first build modules that describe the transmission of thermal radiation through the atmosphere in 4.1 and the interaction of obtained thermal radiation with the imaging sensor in 4.2. These modules enable us to accurately model the thermal imaging process, providing a more robust foundation for our network to learn from.

Finally, we employ an encoder-decoder architecture in our network, which is described in 4.3. This architecture effectively fuses features of the two modules to improve the accuracy of thermal blind road segmentation.

### 4.1. Atmospheric Transmission Module (ATM)

Thermal infrared radiation experiences attenuation as it passes through the atmosphere. This attenuation attributes to two main factors: absorption and scattering. Specifically, carbon dioxide and water vapor present in the atmosphere can absorb thermal infrared radiation, while IR radiation can also be scattered diffusely by particles in the atmosphere. Attenuation of thermal infrared radiation as it travels through air is described by the Bouguer-Lambert-Beer law[30] as Fig. 3 shows. This law provides a relationship between the transmittance of radiance, represented by $T$, and the distance that the radiation has traveled through the air, represented by $d$:

$$T(\lambda, d) = \frac{I(\lambda, d)}{I(\lambda, 0)} = e^{\gamma(\lambda) \cdot d}. \quad (1)$$

Here, $\gamma(\lambda) = n \cdot (\gamma_{abs} + \gamma_{sca})$ represents the total attenuation coefficient, which is primarily composed of two

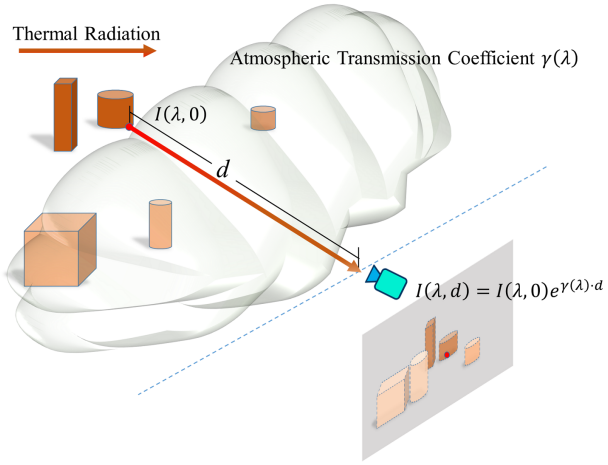


Figure 3. Attenuation of thermal infrared light traveled through the air is primarily composed of absorption and scattering. The transmission is exponentially decreased by the distance $d$ between the object to the camera.

factors: absorption ($\gamma_{abs}$) and scattering ($\gamma_{sca}$). $n$ denotes the volume concentration of the gas.

For an input image $I \in \mathbb{R}^{W \times H \times Channel}$, The output of ATM $I_{ATM}$ is calculated as $I_{ATM} = I \times e^{(\gamma_{abs} + \gamma_{sca}) \times d}$, as shown in Fig. 5(b). Here the $W$, $H$ and $Channel$ denote the width, height and channel of the image, respectively. In our Atmospheric Transmission Module, we set $\gamma_{abs}$, $\gamma_{sca}$ and $d$ as trainable parameters, which are initialized as zeros, zeros and ones, respectively.

The atmospheric transmission and scattering are influenced by the environment. In our method, we adopt the assumption that each pixel represents a shared atmospheric parameter within its specific collection range. Consequently, the ATM module operates on a per-pixel basis in our approach.

The processes here are all in a pixel-to-pixel manner, such as dot multiplications and summations. Thus, $\gamma_{abs}$, $\gamma_{sca}$, $d$ and $I_{ATM}$ are all in a size of $W \times H \times Channel$. This enables the network to effectively learn the attenuation of thermal radiation caused by atmospheric transmission, leading to improved performance in thermal infrared blind road segmentation.

### 4.2. Thermal Inertia Module (TIM)

In Microbolometer Infrared Focal Plane Array (MIFPA), the sensor is constantly exposed to incoming electromagnetic (EM) radiation emitted by the objects in its field of view. As the radiation interacts with the sensing material, it causes a change in the temperature of each individual pixel, which alters its electrical resistance. This electrical resistance is measured by a Read-Out Integrated Circuit (ROIC) at regular intervals to determine the temperature values cor-

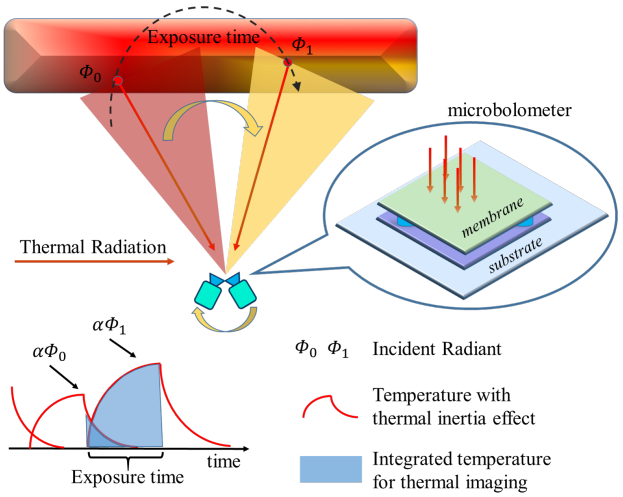


Figure 4. The thermal microbolometer detects incident radiation by measuring the electrical resistance. The incident radiation is captured by the membrane of the microbolometer. A change in electrical resistance between the membrane and the substrate determines the recorded intensity, which in turn is used to generate the final thermal image. This process is impacted by the thermal inertia effect of the microbolometer. This effect can be modeled by parameterizing the thermal conductivity, thermal capacitance and the absorptivity of the microbolometer.

responding to each pixel.

However, if the incoming radiation is continuously changing, the microbolometer is unable to reach thermal equilibrium before the ROIC takes the next measurement. Therefore, the microbolometer must be modeled by taking into account its time-dependent behavior under varying radiation conditions.

Given the absorptance $\alpha(\lambda)$ of the microbolometer, the energy conversion process of the microbolometer can be expressed using the principle of energy conservation as [30, 27]:

$$\alpha\Phi = C_{th}\frac{\mathrm{d}\Delta T}{\mathrm{d}t} + G_{th}\Delta T. \tag{2}$$

Here, $C_{th}$ and $G_{th}$ represent the heat capacitance and heat conductance of the thermal microbolometer, respectively. $G_{th}$ takes into account all of the heat exchange mechanisms, such as conduction, convection, and radiation. In Eq.2, $\Phi$ represents the net radiation power transferred to the detector, while $\alpha\Phi$ represents the net radiant power transferred to the detector which leads to an increase in the temperature, $\Delta T$.

In our module, we set the radiations are square-wave pulses, the temperature difference $\Delta T$ exhibits an exponential rise and decay with a time constant $\tau$. By solving Equation 2, the temperature response $\Delta T(t)$ of the sensor during the rise period is derived as:

$$\Delta T(t) = \frac{\alpha\Phi}{G_{th}}(1 - e^{-\frac{t}{\tau}}), \tag{3}$$

with time constant $\tau = \frac{C_{th}}{G_{th}}$.

To model the effect in a neural network, we assume that the historical radiation are in the decay periods. For the decay period, the temperature response $\Delta T(t)$ of the sensor is given by:

$$\Delta T(t) = \frac{\alpha\Phi}{G_{th}}(e^{-\frac{t}{\tau}}). \tag{4}$$

$\tau$ and $\alpha$ are the time constant and the absorptance of a thermal detector, respectively. The temperature change caused by historical radiation in the last period affects the incoming radiant, resulting in a constant effect on the obtained intensity, which is known as the thermal inertia effect of each pixel.

To model the thermal inertia effect, we consider the historical radiation and the object radiation as $\Phi_0$ and $\Phi_1$, respectively. During the exposure time, the object radiation is in the rise period while the historical radiation is in the decay period. Thus, the obtained intensity can be interpreted as:

$$\begin{aligned} I &= \int_{t_0}^{t_0+t_e}\left[\frac{\alpha\Phi_1}{G_{th}}(1-e^{-\frac{t}{\tau}}) + \frac{\alpha\Phi_0}{G_{th}}e^{-\frac{t}{\tau}}\right]\mathrm{d}t. \\ &= \frac{\alpha}{G_{th}}\Phi_1\left[t_e + \tau e^{-\frac{t_0}{\tau}}(e^{-\frac{t_e}{\tau}}-1)\right] - \frac{\alpha\tau}{G_{th}}\Phi_0 e^{-\frac{t_0}{\tau}}(e^{-\frac{t_e}{\tau}}-1). \end{aligned} \tag{5}$$

Here $t_0$ and $t_e$ represent the starting time in the historical radiation when exposure starts and the duration of exposure to the object radiation, respectively. From Eq. 5 we can derive the object radiation $\Phi_1$ to supply the information for the neural network. The object radiation $\Phi_1$ which is also the output of TIM $I_{TIM}$ is derived as follows:

$$I_{TIM} = \Phi_1 = \frac{I + \frac{\alpha\tau}{G_{th}}\Phi_0 e^{-\frac{t_0}{\tau}}(e^{-\frac{t_e}{\tau}}-1)}{\frac{\alpha}{G_{th}}\left[t_e + \tau e^{-\frac{t_0}{\tau}}(e^{-\frac{t_e}{\tau}}-1)\right]}. \tag{6}$$

Here, $I \in \mathbb{R}^{W\times H\times Channel}$ is the input image. $\alpha$, $C_{th}$, $G_{th}$, $t_0$, $t_e$ and $\Phi_0$ are all trainable parameters initialized with ones in this module. $\tau$ is calculated as $\tau = \frac{C_{th}}{G_{th}}$. Sizes of $\alpha$, $C_{th}$, $G_{th}$, $t_0$, $t_e$, $\tau$ and $\Phi_0$ are all $W \times H \times Channel$.

Thermal mocrobolometers are assembled on a per-pixel basis (each pixel corresponds to a microbolometer). Thus, as shown in Fig. 5(c), the output of TIM $I_{TIM}$ is calculated on a pixel-to-pixel basis.

### 4.3. Architecture

The architecture of TINN is illustrated in Fig. 5(a). Since blind roads for the visually impaired are usually located in fixed positions within cities, we adopt an encoder-decoder architecture primarily based on Transformer blocks

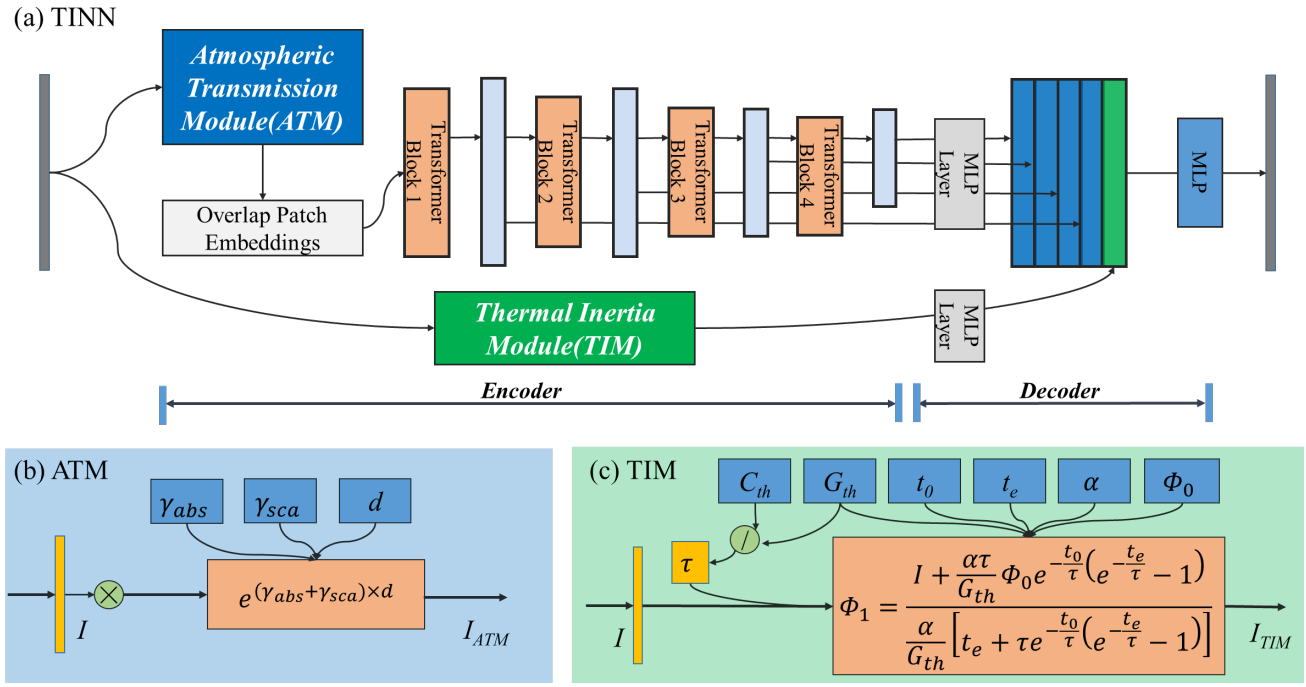


Figure 5. Overview of (a) TINN and the two key blocks: (b) the Atmospheric Transmission Module(ATM) module and (c) the Thermal Inertia Module(TIM). Here the $\gamma_{abs}, \gamma_{sca}, d, C_{th}, G_{th}, t_0, t_e, \alpha, \Phi_0$ and $\tau$ are all trainable parameters. All operations in ATM and TIM are processed on a pixel-to-pixel basis.

to capture global information effectively. To this end, we use SegFormer [34] as our backbone.

The underlying concepts of ATM and TIM rely on a step-by-step progression following a serial structure. Nonetheless, in our network, we present ATM and TIM with discrete and approximated descriptions. Consequently, incorporating a sequential framework like an enhancement module could potentially result in certain inaccuracies. As a result, we have chosen to adopt a parallel framework to ensure both the stability and efficiency of the model, as shown in Fig. 5(a).

**During encoding**, the input image $I$ is passed through the Atmospheric Transmission Module (ATM) to obtain the information combining thermal transmission effect. Specifically, we compute $I_{ATM} = I \times \left(e^{(\gamma_{abs}+\gamma_{sca})\times d}\right)$, where $\times$ represents dot multiplication, $\gamma_{abs}$ and $\gamma_{sca}$ are the absorption and scattering coefficients of the atmosphere, respectively, and $d$ is the distance between the camera and the scene. Afterward, the features go through the 4 hierarchical transformer blocks as suggested by SegFormer. The 4 multi-level features with different sizes are generated as:

$$\langle F_1, F_2, F_3, F_4 \rangle = Enc(I_{ATM}), \tag{7}$$

where $F_1$, $F_2$, $F_3$, $F_4$, and $Enc$ denote the four features and the encoder of SegFormer, respectively.

Additionally, the input image is processed by the Thermal Inertia Module (TIM). In TIM, $\alpha$, $C_{th}$, $G_{th}$, $t_0$, $t_e$, and $\Phi_0$ are all trainable parameters initialized with ones as suggested in 4.2. $\tau$ is calculated by dividing $C_{th}$ and $G_{th}$. The output of TIM $I_{TIM}$ is generated as in Eq. 6.

The operations performed in TIM are all processed on a pixel-to-pixel basis. $I_{TIM}$ is a feature representation of the input image that considers the thermal inertia effect, which is then denoted as $F_5 = I_{TIM}$ in our architecture.

**During decoding**, the first two dimensions of $F_1, F_2, F_3, F_4$ and $F_5$ are resized to the same size, which is 128×128, and the MLP layers transform the last dimension of $F_5$ to 768. As a result, the size of $F_1, F_2, F_3, F_4$ and $F_5$ becomes $128 \times 128 \times 768$. The five features are concatenated and input into the decoder of TINN, which is an MLP recommended by SegFormer, to generate the segmentation mask.

For the segmentation task, we utilize a pixel-wise cross-entropy loss to evaluate individual images. The loss function is defined as:

$$\mathcal{L}_{seg}(\mathbf{P}, \mathbf{Y}) = -\frac{1}{n}\sum^{i}\sum^{j}\mathbf{Y}_{i,j}\log\mathbf{P}_{i,j}, \tag{8}$$

where $\mathbf{P}_{i,j} \in \mathbb{R}$ and $\mathbf{Y}_{i,j} \in \{0, 1\}$ denote the predicted score and groundtruth label of class $j$ at pixel $i$, respectively. $n$ is the total number of pixels in the image.

## 5. Experiments

In this section, we describe the experimental evaluation of our proposed method on our dataset TBRSD in 5.1. We present a comparison of our results both quantitatively and qualitatively with existing methods. Furthermore, we conduct ablation studies to analyze the individual components of our proposed method in 5.2.

### 5.1. Experiment on Thermal Blind Road Segmentation

We conducted an evaluation to assess the effectiveness of our proposed TINN model on the TBRSD dataset. We used the intersection over union ($IoU$) and weighted F-measure ($F_\beta^w$) as the evaluation metrics, as suggested in [20], which are higher the better. TINN is implemented using PyTorch [25] and trained it for 240,000 iterations using stochastic gradient descent with an AdamW optimizer and a weight decay of 0.0001. We used the poly strategy and set the initial learning rate and power to 0.0002 and 0.9, respectively. To initialize TINN, we used a model pre-trained on ImageNet [9]. All input images were resized to $512 \times 512$ for both training and testing, and the final output was resized back to the original input resolution.

To ensure a fair comparison and fully demonstrate the effectiveness of our method, we compared it to 15 state-of-the-art methods, including both visible and thermal semantic segmentation methods. To eliminate the influence of pretraining, all methods were initialized with a model pretrained on ImageNet. Furthermore, we set all methods to the same or recommended training settings for consistency in our evaluation. Among the compared methods, five are based on deep convolution neural networks, including DeepLab V3+ [6], PSPnet [38], UPerNet [33], FCN [18] and SegNet [2]. Meanwhile, five methods, SETR [40], SegFormer [34], Vit-Adapter [7], Mask2Former [8] and MaskDINO [16], are constructed through transformer modules. Additionally, three methods, EC-CNN [15], FTnet[23] and MCNet [35], fall within the domain of thermal semantic segmentation. Notably, Trans4Trans [37] and EOS [19] are both tailored to the needs of visually impaired individuals.

Table 1 shows the quantitative comparison with other methods on TBRSD. The $IoU$ and $F_\beta^w$ are calculated on blind road. Furthermore, we provide the parameter counts for all comparing methods. The count of parameters in our methodology is subject to the selection of the backbone architecture, which remains relatively low in comparison to transformer-based approaches.

Our method surpasses the existing methods with around 1.3% and 0.6% on $IoU$ and $F_\beta^w$. Qualitative comparison with other method is displayed in Fig. 6. Qualitative results show that convolutional neural networks may sometimes ignore the target when the background features are similar to the target. Meanwhile, Transformer-based networks tend to exhibit a higher occurrence of over-segmentation phenomena. Our method obtains more information at a distance and extracts more features in the blurry region, which may be caused by the thermal inertia effect.

Table 1. Results of comparing methods on TBRSD.

| Method | $IoU \uparrow$ | $F_\beta^w \uparrow$ | Params(M) |
|---|---|---|---|
| DeepLab V3+ [6] | 0.8695 | 0.8954 | 62.7 |
| PSPnet [38] | 0.8649 | 0.8926 | 68.1 |
| UPerNet [33] | 0.8616 | 0.8905 | 72.3 |
| FCN [18] | 0.8798 | 0.9145 | 68.6 |
| SegNet [2] | 0.8359 | 0.8654 | 29.5 |
| SETR [40] | 0.8745 | 0.9082 | 318.3 |
| SegFormer [34] | 0.8864 | 0.9376 | 84.7 |
| Vit-Adapter [7] | 0.8796 | 0.9095 | 99.8 |
| Mask2Former [8] | 0.8795 | 0.9126 | 216.0 |
| MaskDINO [16] | 0.8867 | 0.9355 | 223.0 |
| EC-CNN [15] | 0.8642 | 0.8895 | 54.5 |
| MCNet [35] | 0.8556 | 0.8864 | 35.7 |
| FTNet [23] | 0.8426 | 0.8795 | 33.4 |
| Trans4Trans [37] | 0.8525 | 0.8859 | 48.3 |
| EOS [19] | 0.8649 | 0.8925 | 2.36 |
| TINN(ours) | **0.8989** | **0.9439** | 85.3 |

### 5.2. Ablation Studies

In this section, we conducted extensive experiments to evaluate the effectiveness of each proposed component. We evaluated the ablated versions of our method on TBRSD, in order to measure their performance. The detailed results of each ablated version can be found in Table 2.

Table 2. Ablation studies of components of TINN on TBRSD.

| Method | $IoU \uparrow$ | $F_\beta^w \uparrow$ |
|---|---|---|
| Baseline | 0.8864 | 0.9376 |
| Baseline+ATM | 0.8945 | 0.9421 |
| Baseline+TIM | 0.8913 | 0.9405 |
| Baseline+ATM+TIM (TINN) | **0.8989** | **0.9439** |

We set SegFormer as our baseline. The model achieved an $IoU$ of 0.8864 and an $F_\beta^w$ of 0.9376. The ATM component provided a gain of 0.8% $IoU$ and 0.4% $F_\beta^w$, while the TIM component provided a gain of 0.5% $IoU$ and 0.3% $F_\beta^w$. Combining both modules resulted in a gain of 1.3% $IoU$ and 0.6% $F_\beta^w$, which represents the overall performance improvement of the two modules together. To visualize the impact of ATM and TIM on the segmentation task, we apply the the Baseline, Baseline+ATM, Baseline+TIM and Baseline+ATM+TIM methods in Fig. 7.

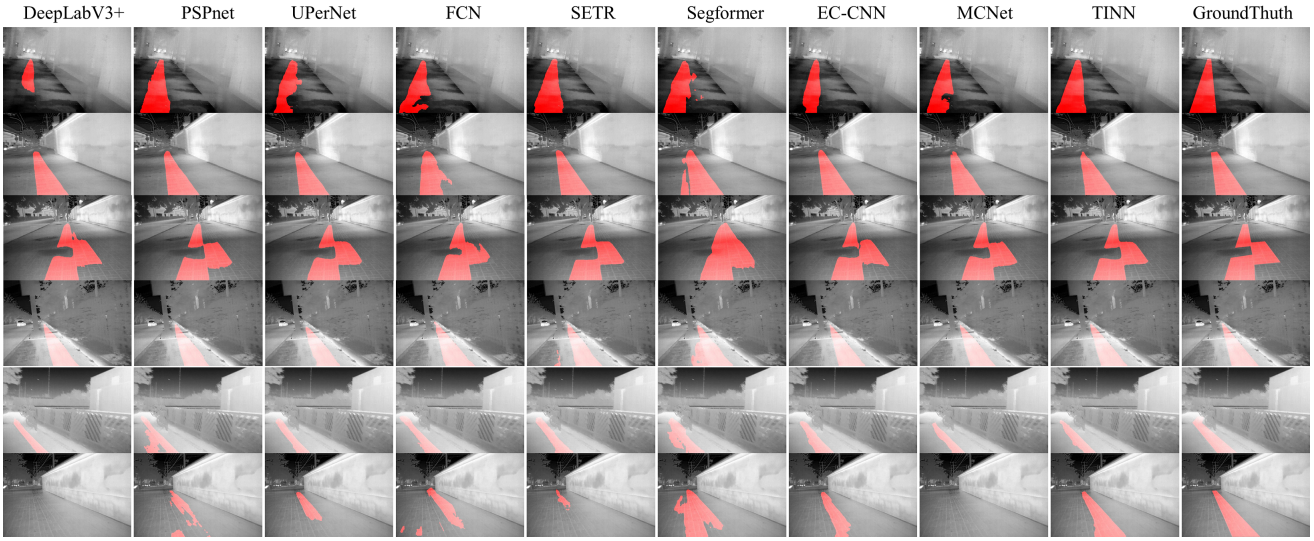


Figure 6. The qualitative comparison between our approach and some existing state-of-the-art methods on TBRSD.

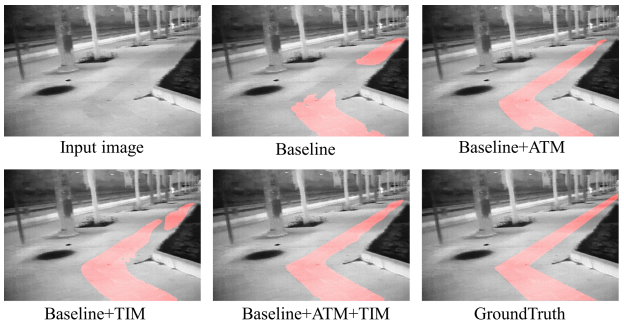


Figure 7. Visual comparison of segmentation results with different methods in ablation studies.

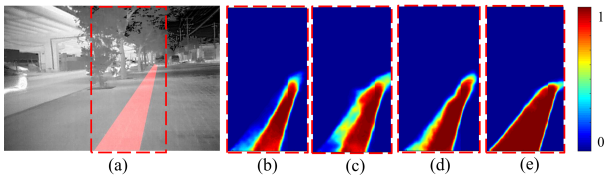


Figure 8. Visual comparison of probabilities of blind road predicted with different methods in ablation studies. (a) Input image with the groundtruth mask, (b) Baseline, (c) Baseline+ATM, (d) Baseline+TIM, (e) Baseline+ATM+TIM.

Furthermore, we visualized the impact of the ablated methods on the predicted probabilities of blind road in Fig. 8. By observing the results, it becomes evident that the presence of distant objects is more pronounced after applying ATM, which effectively mitigates the transmission effect. In another hand, after the implementation of TIM, the probability of localized predictions is enhanced through the reduction of the thermal inertia effect.

## 6. Extended Applications of TINN for Thermal Semantic Segmentation

To further testify the capacity and potential of our method on more complicated thermal segmentation tasks, we perform experiments on two thermal semantic segmentation datasets, MFNet and SODA.

**MFNet** [10] is a dataset that comprises both RGB and IR images. It consists of urban scene images, each annotated with eight categories, namely car, person, bike, curve, car stop, guardrail, color cone, and bump. In this work, we evaluate several methods on the thermal subset of MFNet. Table 3 presents a quantitative comparison of our method with other state-of-the-art methods on MFNet, where we calculate the $IoU$ and $F_\beta^w$ for all categories. Our method outperforms existing methods by approximately 0.6% and 0.5% in terms of $mIoU$ and $F_\beta^w$, respectively.

Table 3. Results of comparing methods on MFNet (thermal subset).

| Method | $mIoU \uparrow$ | $F_\beta^w \uparrow$ |
|---|---|---|
| DeepLab V3+ [6] | 0.4980 | 0.7132 |
| PSPnet [38] | 0.4524 | 0.6941 |
| UPerNet [33] | 0.4856 | 0.7054 |
| SegFormer [34] | 0.5068 | 0.7165 |
| Vit-Adapter [7] | 0.5062 | 0.7106 |
| Mask2Former [8] | 0.5130 | 0.7169 |
| MaskDINO [16] | 0.5103 | 0.7124 |
| EC-CNN [15] | 0.4756 | 0.7106 |
| MCNet [35] | 0.4315 | 0.6945 |
| TINN(ours) | **0.5193** | **0.7226** |

**SODA** [15] contains 2168 real and 5000 synthetically generated thermal images with 21 categories. The real subset is captured by a FLIR camera (SC260), while the synthetic subset consists of thermal images generated from annotated RGB images. In our evaluation, we focus on the real subset and compare the performance of different methods on this subset.

Table 4. Results of comparing methods on SODA(real).

| Method | $mIoU \uparrow$ | $F_\beta^w \uparrow$ |
|---|---|---|
| DeepLab V3+ [6] | 0.6873 | 0.8265 |
| PSPnet [38] | 0.6868 | 0.8247 |
| UPerNet [33] | 0.6745 | 0.8165 |
| SegFormer [34] | 0.6786 | 0.8096 |
| Vit-Adapter [7] | 0.6812 | 0.8199 |
| Mask2Former [8] | 0.6758 | 0.8056 |
| MaskDINO [16] | 0.6632 | 0.7984 |
| EC-CNN [15] | 0.6587 | 0.7965 |
| MCNet [35] | 0.6389 | 0.7846 |
| TINN(ours) | **0.6945** | **0.8356** |

Table 4 presents a quantitative comparison of our method with other state-of-the-art methods on the SODA dataset, where we evaluate the $mIoU$ and $F_\beta^w$ metrics among all categories. Our method outperforms existing approaches by approximately 0.7% and 0.9% in $mIoU$ and $F_\beta^w$, respectively.

The two experiments demonstrate the general capacity of our method for more complicated segmentation tasks such as thermal semantic segmentation tasks.

## 7. Conclusion

In this paper, we propose a novel approach that incorporates thermal imaging into the construction of a segmentation architecture to tackle the thermal segmentation task. Our method models two critical factors in the thermal imaging process, namely atmospheric transmission and thermal inertia, through the atmospheric transmission module (ATM) and the thermal inertia module (TIM), respectively. We then integrate the features generated by ATM and TIM to achieve better segmentation results. We apply our method to blind road segmentation, which is a crucial task in walking assistant systems for visually impaired individuals, where thermal images can improve safety in low-light environments. To evaluate the effectiveness of our approach, we collected a large-scale thermal blind road segmentation dataset, TBRSD, and set a benchmark to assess the quality of our method. The experimental results show that our method achieves promising performance on this task. Additionally, ablation studies demonstrate the impact of ATM and TIM on the segmentation results. Finally, we conducted experiments on the benchmark thermal datasets MFNet and SODA, demonstrating that our method achieves state-of-the-art performance compared to existing methods.

**Limitation and future work** The main limitation of our method is that it is currently designed based on a single backbone. In the future, we can explore incorporating multiple structures for real-world applications. Additionally, our dataset lacks auxiliary information like depth. We can expand our dataset to include additional information in these areas.

**Acknowledgement** This work was supported by the National Natural Science Foundation of China under Grant 62271016, the Beijing Natural Science Foundation under Grant 4222007, and the Fundamental Research Funds for the Central Universities.

## References

[1] José M Álvarez Alvarez and Antonio M López. Road detection based on illuminant invariance. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):184–193, 2010.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[3] Rupert Bourne, Jaimie D Steinmetz, Seth Flaxman, Paul Svitil Briant, Hugh R Taylor, Serge Resnikoff, Robert James Casson, Amir Abdoli, Eman Abu-Gharbieh, Ashkan Afshin, et al. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the global burden of disease study. *The Lancet Global Health*, 9(2):e130–e143, 2021.

[4] Mounir Bousbia-Salah, Maamar Bettayeb, and Allal Larbi. A navigation aid for blind people. *Journal of Intelligent & Robotic Systems*, 64:387–400, 2011.

[5] Zhengcai Cao, Xiaowen Xu, Biao Hu, and MengChu Zhou. Rapid detection of blind roads and crosswalks by using a lightweight semantic segmentation network. *IEEE Transactions on Intelligent Transportation Systems*, 22(10):6188–6197, 2020.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[7] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[10] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.

[11] Marion A Hersh and Michael A Johnson. *Assistive technology for visually impaired and blind people*, volume 1. Springer, 2008.

[12] Lachlan Horne, Jose Alvarez, Chris McCarthy, Mathieu Salzmann, and Nick Barnes. Semantic labeling for prosthetic vision. *Computer Vision and Image Understanding*, 149:113–125, 2016.

[13] Lachlan Horne, Jose M Alvarez, Chris McCarthy, and Nick Barnes. Semantic labelling to aid navigation in prosthetic vision. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3379–3382. IEEE, 2015.

[14] Md Milon Islam, Muhammad Sheikh Sadi, Kamal Z Zamli, and Md Manjur Ahmed. Developing walking assistants for visually impaired people: A review. *IEEE Sensors Journal*, 19(8):2814–2828, 2019.

[15] Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):3069–3082, 2020.

[16] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023.

[17] Yimin Lin, Kai Wang, Wanxin Yi, and Shiguo Lian. Deep learning based wearable assistive system for visually impaired people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision workshops*, pages 0–0, 2019.

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[19] Yinan Ma, Qi Xu, Yue Wang, Jing Wu, Chengnian Long, and Yi-Bing Lin. Eos: An efficient obstacle segmentation for blind guiding. *Future Generation Computer Systems*, 140:117–128, 2023.

[20] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[21] Mehdi Khoshboresh Masouleh and Reza Shah-Hosseini. Development and evaluation of a deep learning model for real-time ground vehicle semantic segmentation from uav-based thermal infrared imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 155:172–186, 2019.

[22] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017.

[23] Karen Panetta, KM Shreyas Kamath, Srijith Rajeev, and Sos S Agaian. Ftnet: Feature transverse network for thermal image semantic segmentation. *IEEE Access*, 9:145212–145227, 2021.

[24] Kibaek Park, Youngtaek Oh, Soomin Ham, Kyungdon Joo, Hyokyoung Kim, Hyoyoung Kum, and In So Kweon. Sideguide: a large-scale sidewalk dataset for guiding impaired people. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10022–10029. IEEE, 2020.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[26] Yuqing Peng. Blind road recognition algorithm based on color and texture information. *Journal of Computer Applications*, 34(12):3585, 2014.

[27] Manikandasriram Srinivasan Ramanagopal, Zixu Zhang, Ram Vasudevan, and Matthew Johnson-Roberson. Pixel-wise motion deblurring of thermal videos. *arXiv preprint arXiv:2006.04973*, 2020.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[29] Z Tang, L Su, C He, and RJ Hu. Research on traffic sign visual recognition technology of guiding robot. *Comput. Technol. Develop.*, 24(9):23–27, 2014.

[30] Michael Vollmer and Klaus-Peter Mollmann. Infrared thermal imaging: Fundamentals, research and applications, 2nd edition, 2017.

[31] Peng Wang and Xiangzhi Bai. Thermal infrared pedestrian segmentation based on conditional gan. *IEEE Transactions on Image Processing*, 28(12):6007–6021, 2019.

[32] Yuanlong Wei and Mincheol Lee. A guide-dog robot system research for the visually impaired. In *2014 IEEE International Conference on Industrial Technology (ICIT)*, pages 800–805. IEEE, 2014.

[33] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[34] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[35] Haitao Xiong, Wenjie Cai, and Qiong Liu. Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene. *Infrared Physics & Technology*, 113:103628, 2021.

[36] He Zhang and Cang Ye. An indoor wayfinding system based on geometric features aided graph slam for the visually impaired. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(9):1592–1604, 2017.

[37] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen. Trans4trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1760–1770, 2021.

[38] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[39] Junwei Zheng, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Materobot: Material recognition in wearable robotics for people with visual impairments. *arXiv preprint arXiv:2302.14595*, 2023.

[40] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021.