# Potential Risk Localization via Weak Labeling out of Blind Spot

Kota Shimomura, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi,
Chubu university
{shimo, hirakawa}@mprg.cs.chubu.ac.jp, {takayoshi, fujiyoshi}@isc.chubu.ac.jp,

## Abstract

*Achieving fully autonomous driving requires not only understanding the current surrounding conditions but also predicting how objects that could lead to potential risks may change in the future. Predicting potential risk regions, especially where pedestrians or vehicles might suddenly appear, is crucial for safe autonomous driving and accident avoidance. Constructing datasets annotated with potential risk regions is costly. Therefore, conventional methods have proposed blind spot estimation using depth maps or segmentation masks through automatic labeling. However, these methods are limited in applicability due to their reliance on camera parameters or point clouds.*

*In this study, we propose a method to automatically generate labels from depth maps and segmentation masks and estimate potential risk regions in 2D. Our automatic labeling algorithm relies solely on images, making it applicable to all onboard camera datasets. To demonstrate the effectiveness of our approach, we define regions where pedestrians or vehicles might emerge from blind spots as potential risk regions and annotate them to create a new dataset extended with potential risk region annotations. Experiments using the Cityscapes Dataset show that weakly training with labels generated by our proposed method achieves equal or superior accuracy compared with supervised training with manually annotated ground truth (GT). Furthermore, experiments using the Mapillary Vistas Dataset and BDD100K Dataset demonstrate the versatility of our approach.*

## 1. Introduction

Advancements in perception, prediction, and planning enabled by deep learning have propelled progress toward achieving fully autonomous driving. Moreover, hardware advancements, such as cameras and LiDAR, are crucial technologies supporting the deployment of large-scale image recognition models. Consequently, real-time processing capabilities during operation have improved, enhancing the ability to avoid traffic risks. However, in autonomous driving and driver-assistance systems, crucial perception functions such as object detection [15, 21] and semantic segmentation [3, 25] are limited to pre-defined objects perceptible to sensors. Consequently, it is challenging to respond to pedestrians or other vehicles suddenly emerging from camera blind spots. Establishing safe driving techniques requires not only understanding the current surrounding conditions but also predicting how objects that could lead to potential risks in the future will evolve. In particular, predicting potential risk regions where pedestrians or vehicles might suddenly appear is crucial for safe autonomous driving and accident avoidance. These regions typically coincide with a driver's blind spots or regions not covered by onboard cameras. Therefore, conventional object detection and semantic segmentation struggle to address these regions because they focus solely on predefined objects and are unable to recognize entirely unobserved regions or objects.

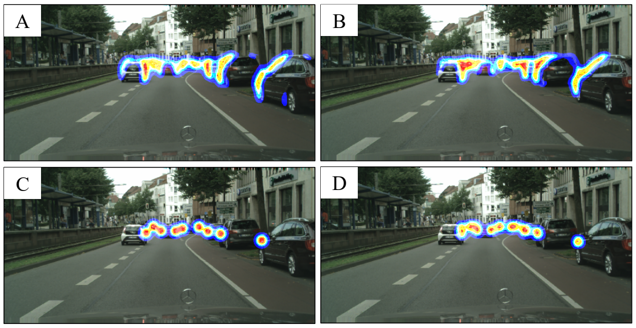

Figure 1. Labels of potential risk regions assigned by different labeling methods and prediction results of the risk region estimation model trained with each label. A is the label of the potential risk region generated by the proposed method, and B is the manually assigned label. C and D are the prediction results of the model trained with labels A and B, respectively.

Why do methods that achieve state-of-the-art accuracy in automotive camera datasets [23, 28, 30] fall short of the risk avoidance required in driving scenarios, even though they are comparable to human recognition capabilities? Therefore, in scenarios such as residential regions with many parked vehicles or in poor visibility conditions like at night or during rain, humans prepare for unexpected incidents by reducing driving speed. Such anticipatory driving represents the crisis avoidance ability inherent in humans and is a crucial capability that deep learning models should acquire to reduce traffic accidents. Methods to obtain crisis avoidance capabilities akin to humans include techniques like ours, which estimate potential risk regions in 2D, as well as methods directly estimating regions in 3D using point clouds and similar data. At this juncture, the most critical aspect is preparing large-scale datasets with annotations of potential risk regions. However, annotating potential risk regions with clear definitions for large datasets is challenging. In particular, annotations on videos or point clouds are practically impossible due to their difficulty and the immense time required.

The main contributions of our study are twofold. First, we introduce an algorithm that automatically generates labels for potential risk regions using a highly accurate pretrained model, along with a newly constructed dataset annotated manually with such regions, specifically tailored for the Cityscapes Dataset [6]. This newly constructed dataset is referred to as the Potential Risk Dataset (PRD). The first contribution, the automatic labeling algorithm, a new weakly supervised potential risk regions estimation method using those labels, and, selectively extracts potential risk regions that may adversely affect a vehicle's driving from blind spots, common in any scene. Furthermore, the depth estimation and segmentation models used in the automatic labeling algorithm are not restricted to specific models, enabling adaptability to various scenes and domains, enabling the selection of appropriate models for any driving scene.

The second contribution, the PRD, serves as a benchmark for the potential risk region estimation task and is used to evaluate the labels generated by the proposed automatic labeling algorithm. Experimental results demonstrate that a potential risk region estimation model trained using labels generated by the automatic labeling algorithm achieves comparable accuracy to a supervised model trained with manually annotated ground truth (GT), as shown in Fig. 1. Additionally, experiments conducted using the Mapillary Vistas Dataset [20] and BDD100K Dataset [27], constructed with images collected from various regions, demonstrate the adaptability of the proposed automatic labeling algorithm across diverse scenes.

In summary, the contributions of this study are as follows.

- Extension of the Cityscapes Dataset through the addition of annotations for potential risk regions.
- A new weakly supervised method for estimating potential risk regions using weakly labels generated by the proposed automatic labeling algorithm.
- Proposal of an automatic labeling algorithm for potential risk regions adaptable to all onboard camera datasets.

## 2. Related Work

Establishing safe driving techniques requires preparation for pedestrians or vehicles emerging from blind spots of drivers or onboard cameras. Therefore, methods for estimating risk regions in advance using sensor information installed in vehicles have been studied. These methods can be broadly classified into two categories. One is the use of fully supervised learning methods, which define risk regions where pedestrians or vehicles may emerge, annotate existing datasets, and train risk region estimation models using fully labeled data. The other is weakly supervised learning methods that utilize various sensors attached to vehicles to detect blind spots within onboard cameras and train blind spot estimation models by treating them as labels. In this section, we review risk region estimation using fully supervised labels and blind spot estimation using weakly supervised labels.

### 2.1. Risk Region Estimation

Several methods have been proposed for estimating risk regions using manual annotations as GT labels, defining regions where pedestrians or vehicles may emerge [17, 22]. Kozuka et al. [17] defined risk regions as those where pedestrians may emerge and proposed a risk region estimation method utilizing one-pixel annotations at the center of risk regions and a regression-based loss function. However, since predefined risk regions are limited to those where pedestrians may emerge, it is challenging to address regions where vehicles or objects other than pedestrians may emerge. Shimomura et al. [22] defined potential risk regions as those where pedestrians or vehicles may emerge and constructed a new dataset by adding annotations to the Cityscapes Dataset. Similar to Kozuka et al., the newly constructed dataset adopts the one-point annotation method, resulting in a small proportion of image regions and limited information available during training. To address the issue of limited information in potential risk regions during training, this study solved the problem by expanding distance-based labels using depth bias in outdoor images [4] and Gaussian filtering. Additionally, the formulation of a potential risk region as a probability distribution estimation problem enabled the introduction of a loss function to handle imbalanced teacher labels, solving the issue of considering relationships between pixels and enabling the estimation of potential risk regions.

### 2.2. Blind Spot Estimation

Instead of manually creating teacher labels for blind spots, methods for estimating blind spots using LiDARs, cameras, and pre-trained depth estimation or segmentation models have also been explored [9, 16, 24, 29]. Zhou et al. [29] proposed a method to identify the nearest blind spot regions by calculating vertical gradients using depth maps and to

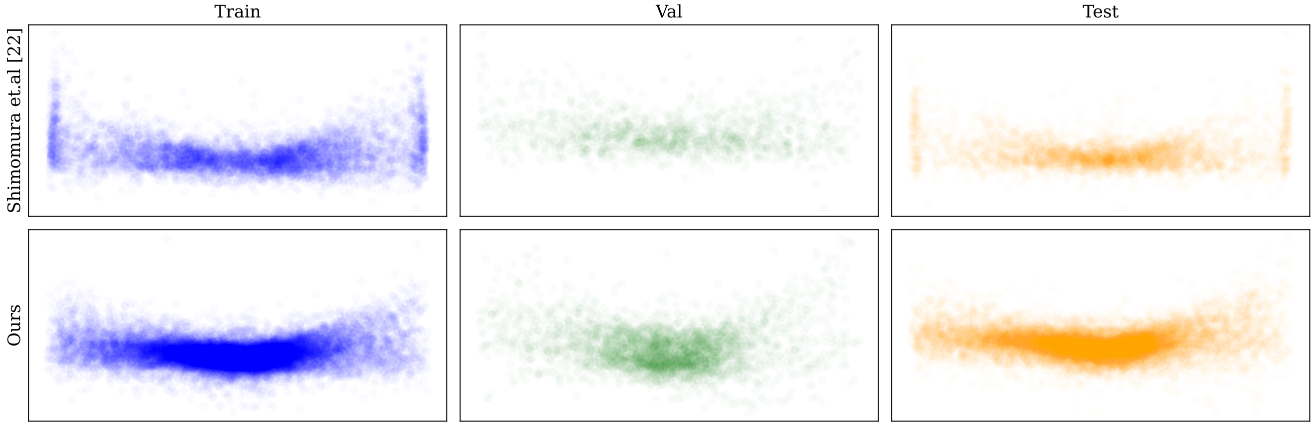

Figure 2. Potential Risk Regions Distribution for Cityscapes Dataset.The x-axis and y-axis axes correspond to Cityscapes' default image size.

early detect pedestrians appearing from blind spots through pedestrian object detection. However, it remains challenging to address multiple blind spot regions present in arbitrary scenes. Furthermore, in early detection, it is difficult to avoid collisions unless the moving speeds of the vehicle and pedestrians emerging from blind spots are slow enough to be recognizable by the camera. Sugiura et al. proposed a method to generate probabilistic multi-hypothesis occupancy grid maps (OGMs) for blind spot regions and obstacles using measurements from 2D range sensors or monocular camera images [24]. While there are several methods to generate OGMs using surrounding vehicle environmental information [1, 13], they face challenges such as the high cost of learning using point cloud data for generating teacher data and limited operable scenes.

Odagiri et al. [16] defined blind spots as the points of contact between the road surface and potential risk objects emerging from them, in contrast to the aforementioned methods that generate OGMs. Therefore, it becomes possible to directly estimate the distance to blind spots from a single depth map without the need for complementary frames for invisible regions like Fukuda et al. [9]. However, since the generated blind spot regions are projected onto the road surface, there is a possibility of adverse effects on driving when the eye level drops. Additionally, the evaluation through experiments is limited by the small number of manually annotated labels provided for the KITTI Dataset [10], indicating insufficient comparison of accuracy with models trained using manually annotated labels and effectiveness assessment using other datasets.

## 3. Potential Risk Dataset

We provide precise annotations of potential risk regions in the Cityscapes Dataset collected in 50 cities, including Germany and neighboring countries.

| Data | Risk scene | UnRisk scene | Workers |
|---|---|---|---|
| Train | 2741 | 234 | 4 |
| Val | 461 | 39 | 3 |
| Test | 1460 | 40 | 4 |

Table 1. Annotation details for Cityscapes Dataset [6], where Risk scene indicates the number of scenes that contain potential risk regions and UnRisk scene indicates the number of scenes that do not contain potential risk regions and UnRisk scene indicate the number of scenes that contain and do not contain potential risk regions.

### 3.1. Potential Risk Annotations

The newly constructed Potential Risk Dataset (PRD) was annotated by four individuals with extensive driving experience. Annotating potential risk regions is challenging due to the inability to define them on the basis of object boundaries, as in object detection or semantic segmentation. Therefore, similar to Kozuka et al. [17], we adopted a 1-point annotation method in this study. Annotators were assigned to each city in Cityscapes Dataset to prevent multiple annotators from annotating the same city. The annotation criteria were defined as regions where pedestrians or vehicles may emerge and affect the vehicle's trajectory. Therefore, not all scenes in the Cityscapes GtFine Dataset [6] are annotated. Conventional potential risk region datasets [22] consider intersections and crossroads where pedestrians or vehicles may emerge but do not affect the vehicle's trajectory. However, annotating potential blind spots caused by intersections and crossroads is challenging because potential ones cannot be accurately annotated from 2D onboard camera images due to the inability to consider road structures. Therefore, annotations for intersections and crossroads are not provided in PRD. Tab. 1 shows the number of annotators for

| Dataset | Train | Val | Test |
|---|---|---|---|
| Shimomura et.al [22] | 6,673 | 1,403 | 3,559 |
| Ours | 20,987 | 3,463 | 13,972 |

Table 2. Number of potential risk regions for each data in Cityscapes[6].

each dataset and the percentage of scenes containing potential risk regions. The annotation for the entire dataset takes 4160 minutes, requiring approximately 54 seconds per image.

### 3.2. Statistical analysis

We compare the number of potential risk region annotations in our newly constructed PRD and the dataset by Shimomura et al. [22] in Tab. 2. It can be observed that there are more than three times as many annotations for potential risk regions in both the Train and Test datasets and more than twice as many annotations in the Val dataset. Furthermore, the distribution of potential risk region annotations per data is illustrated in Fig. 2. Unlike Shimomura et al., our newly constructed dataset does not consider blind spots caused by intersections and crossroads as potential risk regions, resulting in a concentration of risk regions towards the image centers.

## 4. Proposed Method

This section describes main component: an automatic labeling method for discontinuous regions detected using gradients of depth maps used to learn weakly supervised latent risk region estimation.

### 4.1. Automatic labeling algorithm for potential risk regions

**Detection of blind spot regions using depth maps**. As discussed in Section 3, manual annotation of potential risk regions can be costly even when using a single-point annotation method. In our proposed approach for generating potential risk labels using weak supervision, we extract blind spot regions for any object by detecting discontinuities in predicted depth maps and performing class selection using segmentation. An overview of the proposed automatic labeling method for potential risk regions is shown in Fig. 3. The regions where pedestrians or vehicles might suddenly emerge cannot be visually confirmed by drivers and may appear suddenly from behind obstacles. Therefore, in this study, we define blind spot regions calculated from depth maps as potential risk regions. Our approach uses ZoeDepth [2], capable of estimating relative and absolute depth without requiring fine-tuning, for depth estimation. Our discontinuity detection algorithm is adaptable to any scene, and thus differences in depth scale or variations in depth over time at different instances pose no issue. To identify image coordinates where distances change discontinuously using the normalized depth map $\widehat{D}$ predicted by ZoeDepth for any image $I \in \mathbb{R}^{(C\times H\times W)}$, we perform second-order differentiation in the horizontal direction. We define functions $F_{in}$ and $F_{out}$ to segment the horizontal derivatives of the depth map.

$$F_{\text{in}}(\hat{D},x,y) = \begin{cases} \max(\hat{D}_{x+1,y} - \hat{D}_{x,y}, 0) & \text{if } x < \frac{W}{2} \\ \max(\hat{D}_{x-1,y} - \hat{D}_{x,y}, 0) & \text{otherwise,} \end{cases} \tag{1}$$

$$F_{\text{out}}(\hat{D},x,y) = \begin{cases} \max(\hat{D}_{x-1,y} - \hat{D}_{x,y}, 0) & \text{if } x < \frac{W}{2} \\ \max(\hat{D}_{x,y} - \hat{D}_{x+1,y}, 0) & \text{otherwise,} \end{cases} \tag{2}$$

where, $x$ and $y$ denote the indices of the matrix. In computing the gradient between adjacent pixels in $\widehat{D}$ Eq. (1) extracts the depth variations from the left half of the image towards the right and vice versa. Eq. (2) isolates the depth variations from the right half towards the left and vice versa. By utilizing Eq. (1) and Eq. (2), we can compute the horizontal distance gradients $G^{+} \in \mathbb{R}^{W\times H}$ and $G^{-} \in \mathbb{R}^{W\times H}$ from the predicted depth map $\widehat{D}$.

$$G^{+}(x,y) = F_{\text{in}}(-F_{\text{in}}(\hat{D},x,y),x,y), \tag{3}$$

$$G^{-}(x,y) = F_{\text{out}}(-F_{\text{out}}(\hat{D},x,y),x,y), \tag{4}$$

where, Eq. (3) and Eq. (4) invert the magnitude of the first-order differentials to store the gradient computation results in neighboring pixels. Additionally, to define potential risk regions as blind spots where pedestrians or vehicles may suddenly emerge, we use a threshold to confine the risk regions to spaces where vehicles or pedestrians might appear. The final blind spot region is computed as follows.

$$\widehat{G}^{+} = \{(x,y) | G^{+}[x,y] \geq \sigma\}, \tag{5}$$

$$\widehat{G}^{-} = \{(x,y) | G^{-}[x,y] \geq \sigma\}. \tag{6}$$

Here, $\sigma$ represents the threshold. In this study, empirically, we set $\sigma = 0.05$. The horizontal distance gradients obtained through the second-order differentiation, $\widehat{G}^{+}$ and $\widehat{G}^{-}$, are illustrated in Fig. 4.

The $\widehat{G}^{+}$ computed from Eq. (5) detects blind spot regions unrelated to potential risk regions, as illustrated in Fig. 4. Therefore, from the detected blind spot region $\widehat{G}^{+}$ using segmentation, we extract potential risk regions where pedestrians or vehicles may emerge.

**Automatic labeling of potential risk regions using blind spot and segmentation**. In this study, we use mask2former [5] for segmentation, extracting only the discontinuous regions for any class mask to create labels for potential risk regions. For any image $I \in \mathbb{R}^{(C\times H\times W)}$, a binary mask $BM \in \mathbb{R}^{(H\times W)}$ representing vehicles is created from the

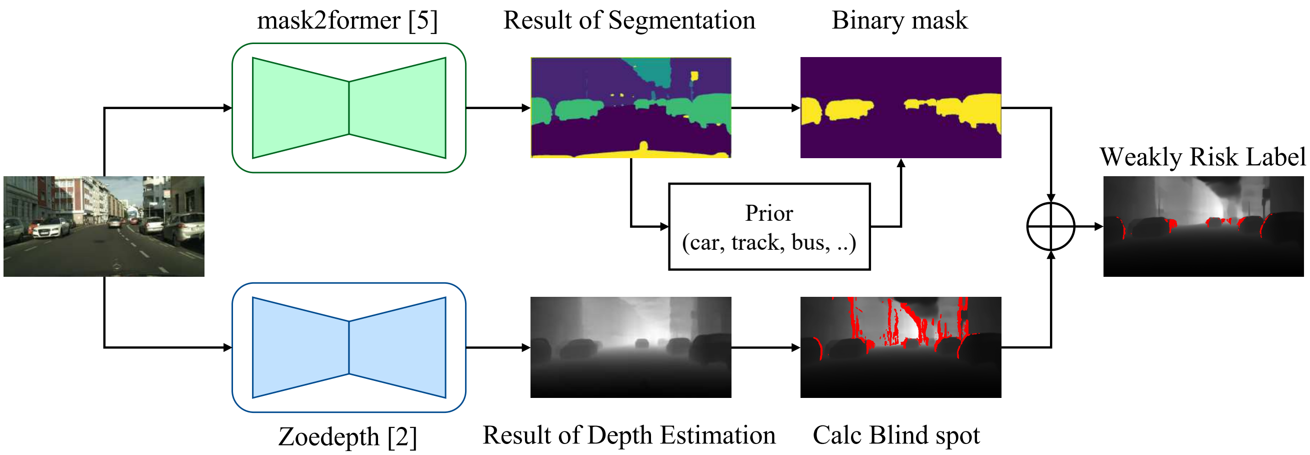

Figure 3. Overview of the automatic labeling method. Blind spots detected by the proposed method are shown in red. Symbols indicate pixel-wise logical AND. The region is intentionally enlarged for visibility.

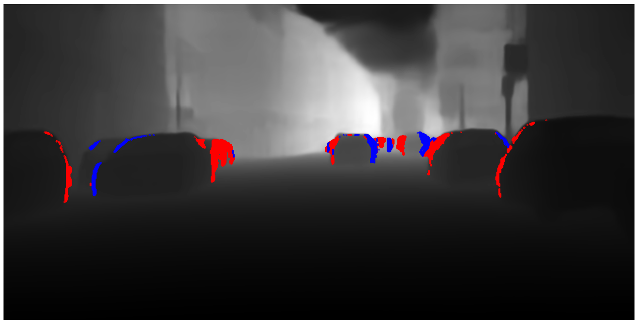

Figure 4. Blind spots caused by the vehicle classes detected by the proposed method. Blue indicates forward blind spots for a given vehicle class, and red indicates rear blind spots.

segmentation mask $M \in \mathbb{R}^{(H \times W)}$ predicted by mask2former [5]. The final label $l_i$ for potential risk regions generated by automatic labeling is calculated as follows.

$$l_i = \widehat{G}^{+} \wedge BM, \tag{7}$$

where, $\wedge$ the pixel-wise logical AND.

**Preprocessing of potential risk region labels using depth.** In previous studies, rather than directly using potential risk regions that occupy a small percentage of the image region for training, preprocessing with outdoor image depth bias [22] was used to address imbalanced labels. However, using preprocessing with outdoor image depth bias makes it difficult to consider accurate depth in the depth direction. Additionally, it is challenging to handle scenes where the road surface on which the vehicle travels does not always extend straight ahead. Therefore, in this study, we expand the labels $l_i$ of potential risk regions at the pixel level using the depth map $\widehat{D}$ used during automatic labeling. The final calculation of potential risk region labels is performed by the Algorithm 1. In this study, we set scale-f=30.

**Algorithm 1** Preprocess

```
Require: binary matrix $\mathcal{BM}$, depth map $\mathcal{D}$, scale-f $\mathcal{S}$
$\mathcal{EM}(expansion\_matrix) \leftarrow$ copy($\mathcal{BM}$)
$\mathcal{RM}(radius\_matrix) \leftarrow$ max($\mathcal{D}$) $- \mathcal{D}$
$\mathcal{MR}(max\_radius) \leftarrow$ max($\mathcal{EM}$) $\times \mathcal{S}$
$output \leftarrow$ zeros((all pixels))
for $y, x \in$ range(all pixels) do
    if $\mathcal{BM}_{(}y, x) = 1$ then
        for $i, j \in$ range(nearest pixels) do
            Nearest pixels range is $(-\mathcal{MR}, \mathcal{MR} + 1)$)
            if $i^2 + j^2 \leq \mathcal{MR}^2$ then $(y', x') \leftarrow (y+i, x+j)$
                if $0 \leq y' < img\_h \wedge 0 \leq x' < img\_w$ then $output_{(}y', x') = 1$
                end if
            end if
        end for
    end if
end for
$output$ : Potential Risk Regions after Preprocess
```

# 5. Experiments

## 5.1. Datasets

For evaluation, we adopted two in-vehicle camera datasets. To investigate the effectiveness of the labels generated by automatic labeling and demonstrate the adaptability of our proposed method to various scenes, we conducted experiments using Cityscapes annotated manually by us [6] and Mapilarry Vistas [20]. These datasets are detailed as follows.

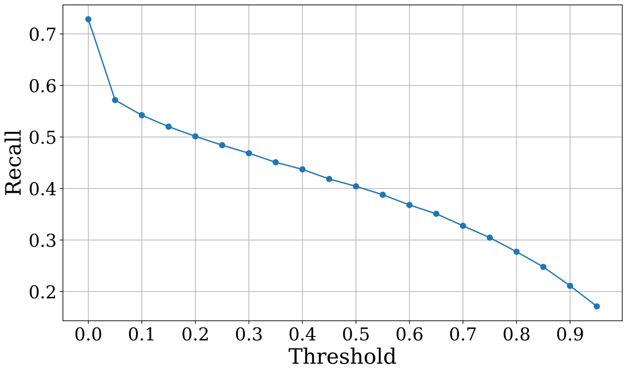

Figure 5. Variation of Recall with respect to the threshold of the weakly label after the application of Gaussian filtering.

| CC | Recall | mAR |
|---|---|---|
| 0.4982 | 0.7284 | 0.4321 |

Table 3. Justification of weakly labels in the Cityscape testset.

**Cityscapes GtFine [6]**: Consists of 5,000 images with precise segmentation annotations collected from 50 cities in Germany. The training and evaluation datasets contain 2,975 and 500 images, respectively, while the remaining 1,525 images are for testing.

**Mapillary Vistas Dataset [20]**: Comprises 25,000 images with precise segmentation annotations collected from various regions worldwide. The training and evaluation datasets contain 18,000 and 2,000 images, respectively, while the remaining 5000 images are for testing.

**BDD100K Dataset [27]**: Comprises 100,000 images with precise segmentation annotations collected from various regions US. The training and evaluation datasets contain 70,000 and 10,000 images, respectively, while the remaining 20,000 images are for testing.

### 5.2. Implement Details

**Automatic Label Generation:** In Cityscapes GtFine [6], Depth maps are generated using ZoeDepth [2] with image sizes of 1024 × 2048. Segmentation masks for class selection are created with image sizes of 512 × 1024 using mask2former [5]. During automatic labeling, the matrix size is adjusted using inter-nearest [11] to ensure that the depth map and segmentation mask have the same size. The final potential risk regions are generated from the occluded regions on the far side of any instance on the basis of the prior knowledge learned from the depth map and mask2former [5].

In Mapillary Vistas Dataset [20] and BDD100K Dataset [27], Depth maps are generated with image sizes of 774 × 1032 using ZoeDepth [2]. Similarly, segmentation masks for class selection are created with the same image size as the depth map using mask2former [5]. Subsequent automatic labeling follows the same procedure as that for Cityscapes. Hereafter, we refer to the potential risk regions generated by the proposed method as weakly labels.

**Network Structure:** The potential risk regions represent blind spots caused by vehicles or obstacles captured by cameras or human gaze. Therefore, it is crucial for the network estimating potential risk regions to mimic human visual processing mechanisms. To address this, we adopt TranSalNet proposed by Lou et al. [19], focusing on research in visual saliency prediction [8, 14, 26]. Lou et al. noted that convolutional neural network (CNN) architectures tend to lose distant contextual information in extracted image features due to CNN-specific inductive biases [7]. Thus, TranSalNet integrates Transformer components into a CNN to consider the ability of the human visual system to understand local and global visual information. Moreover, in our experiments, we use Resnet-50 [12] as the backbone network.

**Training Details:** Training is conducted using Nvidia RTX A6000. Subsequently, a Gaussian filter is applied with a kernel size of 5 × 5. Images are resized using inter-nearest interpolation to 320 × 640 for Cityscapes and 288 × 384 for Mapillary Vistas. The loss function uses a weighted linear combination of Exponentially Weighted MSE Loss and Total Variation Distance, similar to Shimomura et al. [22]. We set the initial learning rate to $10^{-4}$ using the AdamW algorithm [18], and the model is trained for 60 epochs with a batch size of 16.

### 5.3. Justification of Weakly labels

The primary goal of estimating potential risk regions are to reduce the number of traffic accidents due to undetected risks when considering operation in real-world applications. Therefore, the validity of weakly labels is evaluated by the Correlation Coefficient (CC) between labels and Average Recall (AR). Tab. 3 presents the quantitative evaluation results using the Cityscapes test data. Recalls for each threshold are shown in the Fig. 5. The correlation coefficients and Recall indicate that the weakly label generated by the proposed method is comparable to the manually annotated ground truth.

### 5.4. Evaluation metrics for predicting potential risk regions

To evaluate the consistency between manually annotated potential risk region labels, labels generated by the proposed method, and predicted potential risk region labels, both location-based and distribution-based evaluation metrics can be considered. Distribution-based evaluation metrics like Correlation Coefficient (CC), adopted in previous studies [22], do not penalize for undetected or falsely detected regions, which is deemed inappropriate for tasks like ours that prioritize undetected or falsely detected regions.

| Train Label | AUC | Precision | Recall | F1-score | Specificity | F2-score |
|---|---|---|---|---|---|---|
| Supervised | 0.6634 | **0.9104** | 0.3283 | 0.4819 | **0.9985** | 0.3762 |
| Weakly | **0.6890** | 0.7103 | **0.3857** | **0.4975** | 0.9924 | **0.4235** |

Table 4. Quantitative results with the Cityscapes Test dataset. The evaluation was performed with manually assigned GT labels of potential risk regions.

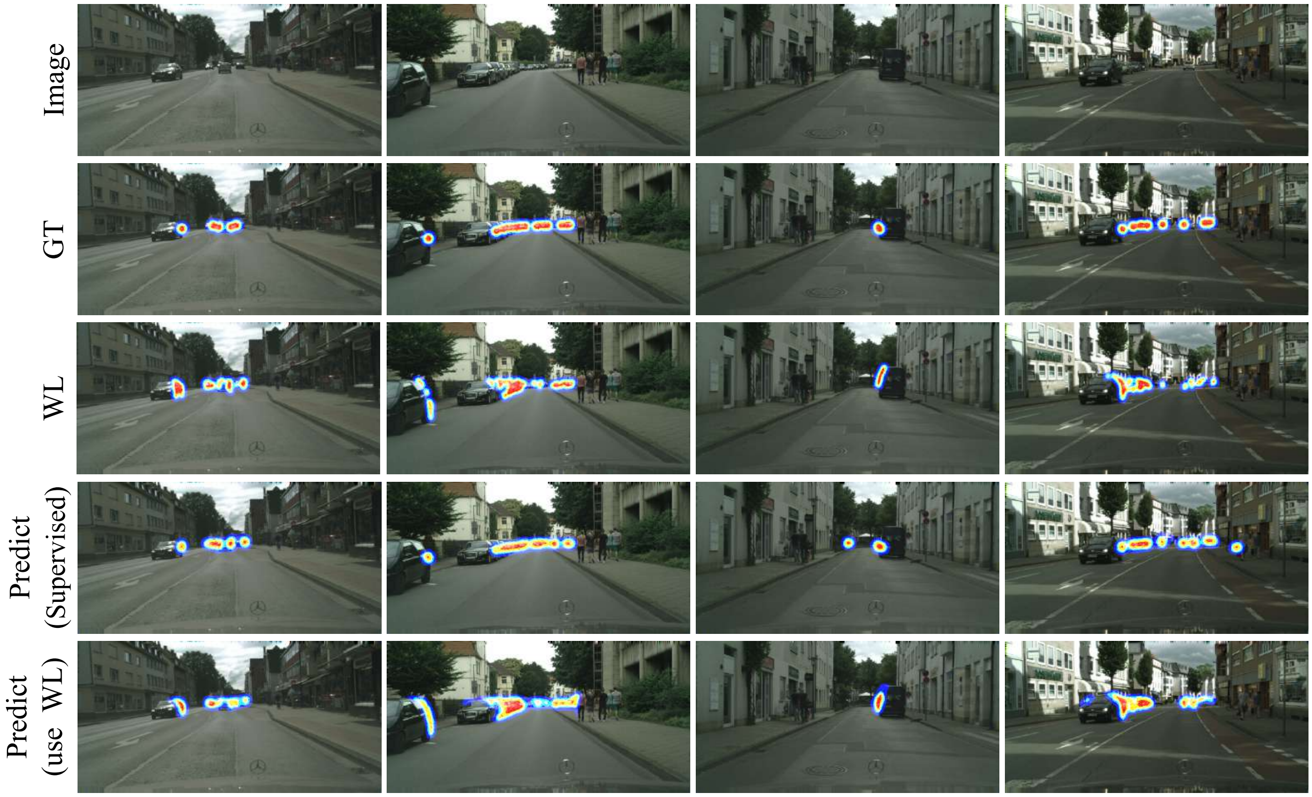

Figure 6. Visualization results from the Cityscapes Test, where GT is ground truth and WL is Weakly Label. The fourth row shows the prediction results for the model trained with Weakly Label, and the fifth row shows the prediction results for the model trained with GT labels.

In this study, we focus on the accuracy of detected regions of potential risk regions and adopt location-based evaluation metrics such as area under the curve (AUC), Precision, Recall, F1-score, and Specificity. Additionally, considering the significance of undetected potential risk regions as crucial incidents leading to traffic accidents, we also incorporate the F2-score, which prioritizes Recall. The F2-score is calculated as follows in Eq. (8).

$$\text{F2-score} = \frac{5 \cdot \text{Precision} \times \text{Recall}}{4 \cdot \text{Precision} + \text{Recall}} \tag{8}$$

### 5.5. Comparison with fully supervised learning

We compare the accuracy of models trained using manually annotated GT labels and potential risk region labels generated by the proposed method using the Cityscapes Dataset. Evaluation is conducted using the provided GT annotations. We aim to achieve equal or superior accuracy compared with fully supervised learning with GT labels, using potential risk region labels generated by our automatic labeling algorithm, which does not require extensive annotation costs. Tab. 4 presents the quantitative evaluation results using the Cityscapes test data. As the results indicate, models trained using weakly labeled potential risk regions generated by our proposed automatic labeling algorithm demonstrate equal or superior accuracy compared with fully supervised models. We observed improvements in AUC, Recall, F1-score, and F2-score compared with fully supervised learning. Specifically, AUC improved by 0.0256 pt, Recall by 0.057 pt, and F2-score by 0.047 pt. These findings demonstrate that weakly labeled potential risk regions generated by our proposed automatic labeling algorithm are effective for training

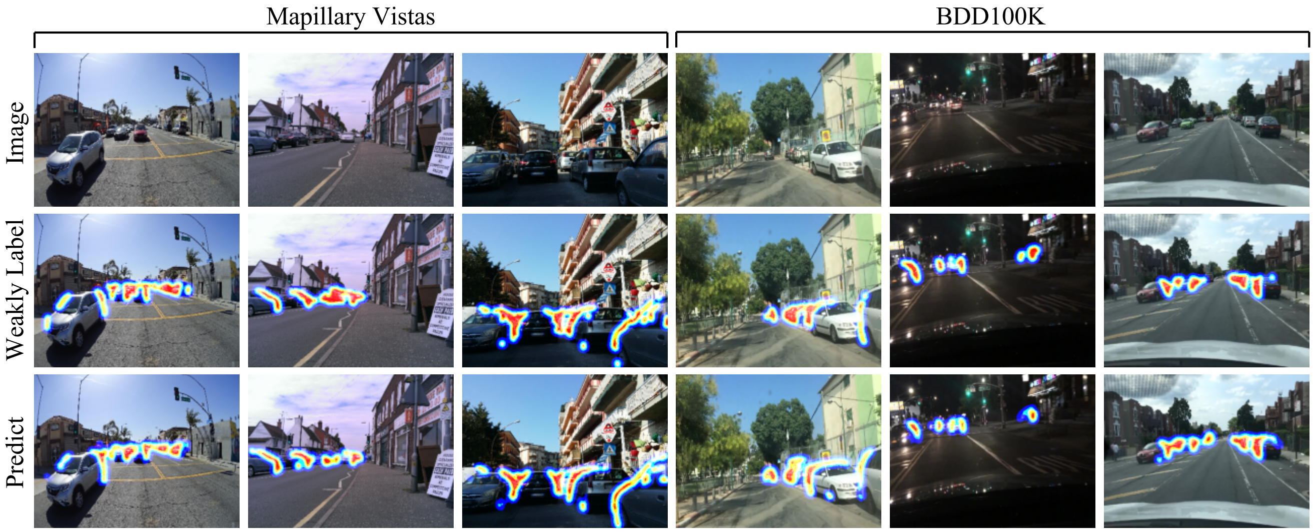

Figure 7. Visualization results from the Mapillary Vistas Test and BDD100K Test, where the WL (Weakly Label) in the second row is generated by the proposed automatic labeling algorithm.

potential risk region estimation networks, achieving equal or superior effectiveness to the GT labels.

### 5.6. Model Visualization

Fig. 6 illustrates qualitative evaluations on the Cityscapes test data. The weakly labels generated by the proposed method are acceptable potential risk region labels compared with the GT labels. Moreover, focusing on the third row, the proposed method can generate accurate potential risk region labels even for distant objects, where manual annotation would be challenging. From the results, the proposed method is not limited to specific regions or scenes. Overall, the proposed method demonstrates the ability to detect potential blind spots behind arbitrary objects in various scenes, indicating that training of potential risk region estimation models is possible on any data as long as depth estimation and semantic segmentation are applicable.

### 5.7. Versatility of the proposed method

The proposed method is not limited to specific datasets. Therefore, we demonstrate the high versatility of the proposed method using the Mapillary Vistas Dataset and The BDD100k Dataset [27]. Since both datasets lack ground truth annotations, we discuss qualitatively based on the evaluation results shown in Figure 7. The left three columns depict results using the Mappillary Vistas dataset, while the right three columns show results using the BDD100K dataset. From Figure 7, it is evident that the weakly labels generated by our proposed method indicate potential risk regions where pedestrians or vehicles might emerge, not only in the Cityscapes dataset but also in the Mappillary Vistas and BDD100K datasets. Furthermore, the predictions of the models trained using weakly labels demonstrate the ability to predict potential risk regions with high accuracy qualitatively.

While our evaluation is currently limited to qualitative assessment, we plan to annotate potential risk regions for Mappillary Vistas and BDD100K datasets in the future, expanding our evaluation beyond qualitative assessment.

## 6. Conclusion

To achieve potential risk region estimation not limited to specific domains, we constructed a new dataset along with an automatic labeling algorithm that does not rely on camera parameters or point clouds for evaluation. The quantitative evaluation experiments, comparing models trained with manually annotated GT labels and models trained with automatically generated weakly labels, are, to our knowledge, the first of their kind. The automatic labeling algorithm demonstrated its capability to generate labels equivalent to manual annotations by focusing on blind spot regions behind arbitrary instances. Additionally, we trained models using the GT labels from the newly constructed potential risk region dataset and weakly labels generated by our proposed method to investigate the effectiveness of our approach. The results showed that models trained using weakly labels generated by our proposed automatic labeling algorithm achieved equal or superior accuracy compared with models trained using GT labels. This underscores the effectiveness of our approach in potential risk region estimation.

## References

[1] Oladapo Afolabi, Katherine Driggs–Campbell, Roy Dong, Mykel J. Kochenderfer, and S. Shankar Sastry. People as sen-

sors: Imputing maps from human actions. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2342–2348, 2018. 3

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 4, 6

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1

[4] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 2

[5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 4, 5, 6

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 4, 5, 6

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 6

[8] Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified Image and Video Saliency Modeling. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 2020. 6

[9] Taichi Fukuda, Kotaro Hasegawa, Shinya Ishizaki, Shohei Nobuhara, and Ko Nishino. Blindspotnet: Seeing where we cannot see. In *Proc. of European Conference on Computer Vision (ECCV) Workshops – Autonomous Vehicle Vision Workshop*, 2022. 2, 3

[10] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[11] Pascal Getreuer. Linear Methods for Image Interpolation. *Image Processing On Line*, 1:238–259, 2011. https://doi.org/10.5201/ipol.2011.g_lmii. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[13] Masha Itkina, Ye-Ji Mun, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. Multi-agent variational occlusion inference using people as sensors. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4585–4591, 2022. 3

[14] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction, 2021. 6

[15] M. Karthi, V Muthulakshmi, R Priscilla, P Praveen, and K Vanisri. Evolution of yolo-v5 algorithm for object detection: Automated detection of library books and performace validation of dataset. In *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–6, 2021. 1

[16] Odagiri Kazuya and Onoguchi Kazunori. Monocular blind spot estimation with occupancy grid mapping. In *International Conference on Machine Vision and Applications, MVA*, pages 1–6. IEEE, 2023. 2, 3

[17] Kazuki Kozuka and Juan Carlos Niebles. Risky region localization with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017. 2, 3

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 6

[19] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 2022. 6

[20] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5000–5009, 2017. 2, 5, 6

[21] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8, 2023. 1

[22] Kota Shimomura, Hiroki Adachi, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, Masamitsu Tsuchiya, and Yuji Yasui. Potential risk estimation with single monocular camera. In *Secure and Safe Autonomous Driving Workshop and Challenge on CVPR 2023*, 2023. 2, 3, 4, 5, 6

[23] Yosuke Shinya. USB: Universal-scale object detection benchmark. In *British Machine Vision Conference (BMVC)*, 2022. 1

[24] Takayuki Sugiura and Tomoki Watanabe. Probable multi-hypothesis blind spot estimation for driving risk prediction. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 4295–4302, 2019. 2, 3

[25] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021. 1

[26] Sheng Yang, Guosheng Lin, Qiuping Jiang, and Weisi Lin. A dilated inception network for visual saliency prediction. *IEEE Transactions on Multimedia*, 22(8):2163–2176, 2019. 6

[27] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6, 8

[28] Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu. A comprehensive survey on segment anything model for vision and beyond. *arXiv:2305.08196*, 2023. 1

[29] Jiacheng Zhou, Masahiro Hirano, and Yuji Yamakawa. High-speed recognition of pedestrians out of blind spot with pre-detection of potentially dangerous regions. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 945–950, 2022. 2

[30] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1