DIS-Mine: Instance Segmentation for Disaster-Awareness in Poor-Light Condition in Underground Mines

Mizanur Rahman Jewel[†], Mohamed Elmahallawy[‡], Sanjay Madria[†], Samuel Frimpong[§]
[†]Computer Science Department, Missouri University of Science and Technology, Rolla, MO 65401, USA
[‡] School of Engineering & Applied Sciences, Washington State University, Richland, WA 99354, USA
[§]Explosive & Mining Engineering Department, Missouri University of Science and Technology, Rolla, MO 65401, USA
Emails: mj9vc@mst.edu, mohamed.elmahallawy@wsu.edu, madrias@mst.edu, frimpong@mst.edu

***Abstract*—Detecting disasters in underground mining, such as explosions and structural damage, has been a persistent challenge over the years. This problem is compounded for first responders, who often have no clear information about the extent or nature of the damage within the mine. The poor light or even total darkness inside the mines makes rescue efforts incredibly difficult, leading to a tragic loss of life. In this paper, we propose a novel instance segmentation method called DIS-Mine, specifically designed to identify disaster-affected areas within underground mines under low-light or poor visibility conditions, aiding first responders in rescue efforts. DIS-Mine is capable of detecting objects in images, even in complete darkness, by addressing challenges such as high noise, color distortions, and reduced contrast. The key innovations of DIS-Mine are built upon four core components: i) *Image brightness improvement*, ii) *Instance segmentation with segment anything model (SAM) integration*, iii) *Mask R-CNN-based segmentation*, and iv) *Mask alignment with feature matching*. On top of that, we have collected real-world images from an experimental underground mine, introducing a new dataset named ImageMine, specifically gathered in low-visibility conditions. This dataset serves to validate the performance of DIS-Mine in realistic, challenging environments. Our comprehensive experiments on the ImageMine dataset, as well as on various other datasets demonstrate that DIS-Mine achieves a superior F1 score of 86% and mIoU of 72%, outperforming state-of-the-art instance segmentation methods, with at least 15x improvement and up to 80% higher precision in object detection.**

We have made our dataset publicly accessible through ImageMine Dataset.

***Index Terms*—Disaster detection, underground mining, instance segmentation, image enhancement, automatic annotation**

Fig. 1: Samples from our underground mine dataset captured under extremely low-light conditions.

I. Introduction

Underground mining operations are increasingly hazardous as extraction processes delve deeper into the Earth. As mining extends, increased pressures can lead to significant dangers, such as the "crushing" of mine walls and support structures (pillars), or even surface collapse events [1, 2]. In addition, disasters like fires, explosions, or gas and water inundations not only challenge the operational safety of mines but also place miners' lives in jeopardy, requiring immediate, well-coordinated responses [3]. It becomes even more difficult for first responders to enter the mine and assist trapped miners [4]. Limited information about the type and extent of damage inside the mine can lead to delays and increased casualties [5]. The safety of mine workers depends on various interrelated factors, including real-time environmental awareness, hazard identification, communication, training, and experience, all of which are critical during emergencies [6]. For emergency operations to be effective, responsible individuals/first responders must be prepared to manage and delegate essential tasks.

One potential solution is to develop an advanced image segmentation scheme aimed at partitioning an image into coherent regions or segments. This approach could provide first responders with critical information during disasters in mines, helping them identify safe paths and locate miners in need of assistance. Traditional image segmentation methods, such as fully convolutional networks (FCNs) [7], U-Net [8], and DeepLab [9], are capable of achieving high accuracy in segmenting images into different categories or classes. However, these methods treat all objects of the same category uniformly, labeling them with the same class, which poses a significant limitation in our context.

This limitation can be addressed by using instance segmentation, which detects and delineates each distinct object within an image, surpassing the semantic segmentation limitations. Instance segmentation not only identifies the class of an object but also precisely localizes each object instance at the pixel level. However, current instance segmentation approaches share a critical drawback: they perform well *only* in segmenting different categories in images captured in "daylight" or "well-lit" environments but struggle with "low-visibility" or "dark images". The presence of diminished light levels

frequently distorts colors and reduces the contrast between backgrounds and objects. This issue is particularly problematic in underground mining scenarios,*especially during disasters, where the environment is often dark, and existing methods fail to detect anything within these conditions.*

Although some initial steps have been taken to address these challenges in instance segmentation studies [10–12], the results achieved in very dark environments remain suboptimal, with models sometimes achieving less than 50% F1-score on simple datasets. *This leaves a significant open question: is there an instance segmentation approach capable of categorizing objects within images captured in extremely dark environments with acceptable accuracy?*

Contribution. In response to the aforementioned challenges, this paper proposes a cutting-edge instance segmentation model, named DIS-Mine, capable of accurately identifying objects within images captured under poor-light conditions, including complete darkness. To enhance the model's performance, we also introduce an image enhancement technique specifically designed to improve the quality of dark images, thereby enabling more accurate predictions. To validate the proposed techniques, we have collected a real-world dataset from an experimental mine and manually annotated a subset of images into five classes: road, wall, roof, people, and corridor within the underground mine. Additionally, we develop an automatic annotation pipeline that can be used to label the remaining images, as well as new images from similar environments. In short, we make the following contributions:

(1) We propose a novel instance segmentation approach, DIS-Mine, designed to accurately detect and segment each class in images *even under poor-light or near-dark conditions*. This approach is applied to the real-world problem of disaster detection in underground mines, assisting first responders in their rescue efforts.
(2) DIS-Mine features four key innovations: i) *image brightness improvement component* that enhances the ability to distinguish between foreground and background objects in poor-light images; ii) *Instance segmentation with SAM integration component* that applies the enhanced images to an optimized Segment Anything Model (SAM) for precise instance segmentation; iii) *Mask R-CNN-based segmentation component* that utilizes an optimized version of the Mask R-CNN model on the enhanced images to further enhance segmentation accuracy; iv) *Mask alignment with feature matching* that fuses and aligns the outputs from both the optimized SAM and Mask R-CNN models for superior instance segmentation results.
(3) We collect a real-world image dataset, named **ImageMine**, captured in extremely poor-light conditions from an experimental underground mine (see samples in Fig. 1). A small subset of these images was manually annotated into six categories: roads, walls, roofs, people, equipment, and corridors—while we also developed an automatic annotation system to label the remaining images.
(4) Our experimental evaluation demonstrates that DIS-Mine significantly outperforms state-of-the-art approaches across various datasets, including our collected ImageMine dataset. It achieves an F1 score of 86.00% and an mIoU of 72.00%, surpassing existing instance segmentation methods by a substantial margin.

II. Related Work

Instance segmentation task in computer vision aims to identify and delineate each object in an image at the pixel level. Unlike object detection, which provides only bounding boxes around objects, or semantic segmentation, which classifies each pixel into a category without distinguishing between instances of the same category, instance segmentation addresses both localization and classification challenges. Instance segmentation approaches primarily relied on traditional machine learning (ML) techniques. Despite these difficulties, recent efforts have begun to make initial strides in adapting instance segmentation techniques to such challenging conditions.

An approach by Arnab et al. [13] employed per-pixel unary classifiers combined with conditional random fields to maintain spatial consistency. Although these methods were innovative at the time, they faced challenges with the complexity and variability of real-world images. The advent of deep learning (DL) has since revolutionized instance segmentation, a crucial computer vision task that involves outlining individual objects with pixel-level masks. Convolutional neural networks (CNNs) have become the foundation for many state-of-the-art techniques. Among these, Mask R-CNN, introduced by He et al. [14], stands out as one of the most influential models. Mask R-CNN extends the faster R-CNN [15] object detection framework by adding an additional branch to predict segmentation masks for each region of interest (ROI). This advancement has significantly enhanced accuracy and established a new benchmark for subsequent research in instance segmentation. Another significant advancement in instance segmentation is the hybrid task cascade (HTC) [16]. HTC enhances accuracy by integrating object detection and semantic segmentation through a multi-stage architecture. Each stage in HTC refines the predictions from the previous one, resulting in more precise and detailed segmentation masks, especially in complex scenes with overlapping objects.

Real-time instance segmentation has also seen considerable progress. Models such as YOLACT [17] and INSTA-YOLO [18] have been developed to strike a balance between accuracy and speed. YOLACT decomposes the instance segmentation task into parallel sub-tasks, allowing for faster inference times. INSTA-YOLO builds on the YOLO (You Only Look Once) framework, optimizing it for instance segmentation while maintaining real-time performance. These models are vital for applications requiring immediate feedback, such as autonomous driving and robotics.

The development of large-scale datasets has been crucial for advancing instance segmentation. Datasets like COCO (Common Objects in Context) [19] and Cityscapes [20] provide diverse and challenging benchmarks for evaluating model performance. They encompass a wide variety of objects and scenes, enabling researchers to develop and test models

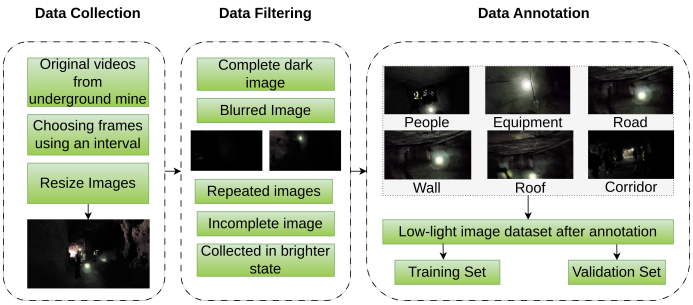


Fig. 2: Construction process for our ImageMine dataset.

under various conditions. Recent research has also explored transformer-based architectures for instance segmentation. The detection transformer (DETR) [21] leverages transformers to model relationships between objects in an image, offering a unified approach to panoptic segmentation. This method has demonstrated promising results, particularly in handling complex scenes with multiple interacting objects.

Despite significant advancements in instance segmentation, challenges remain, such as handling occlusion, varying object scales, and high computational demands. Most importantly, instance segmentation models often perform poorly under low-light conditions, where reduced visibility and increased noise hinder their accuracy. Traditional methods often struggle in such scenarios due to reduced visibility and increased noise, which introduce high-frequency disturbances to neural network feature maps and significantly degrade accuracy. Although some recent works [10, 22] have introduced techniques to enhance accuracy by addressing noise in low-light images, the research into instance segmentation in dark or low-illumination environments remains limited. This challenge is especially pronounced in extreme low-light environments like underground mining, where poor visibility severely complicate object detection. *In this paper, we propose an innovative solution to these challenges, validated on our dataset collected from extremely dark conditions in an underground mine.*

III. ImageMine Dataset Construction

This section details the construction of our underground mining image dataset, ImageMine, which is divided into three main phases: Data Collection, Data Filtering, and Data Annotation. Fig. 2 provides an illustration, with each phase explained in detail in the subsequent subsections.

A. Data collection

The underground surveillance videos were collected from an experimental mine located in Rolla, Missouri, USA, *under extremely low-light conditions*. From each video, about 100-80 images were extracted from the initial frames for manual annotation, focusing on different target objects. To ensure privacy, all data collected complied with consent regulations, ensuring that no facial information or identifiable features of individual miners were captured.

The image acquisition system utilized a high-resolution surveillance camera, featuring a 12 MP wide-angle lens with a variable aperture of f/1.5 to f/2.4 and a 26mm focal length. The camera provided a 77° field of view and captured images at a maximum resolution of 12 megapixels, supporting video recording in up to 4K ultra-high definition (UHD) at 60 frames per second (FPS), using formats like MP4 and HEVC. This advanced system enabled high-quality image capture even under the poor lighting conditions in underground mines.

The collected ImageMine dataset was annotated into **six** categories: road, wall, roof, people, equipment, and corridor. These labels are selected based on the importance of the situation while conducting a rescue operation in an underground mine. Due to poor lighting, some frames lacked target objects such as equipment, people, or corridors, and these frames were excluded. The remaining annotated images were sorted by category to form the final ImageMine dataset.

B. Data Filtering

The original image source of the ImageMine dataset is screened to ensure high quality. This dataset primarily includes images of road, wall, roof, people, equipment, and corridor. However, some images may lack targets, have incomplete targets, or be of poor quality, making identification challenging. For example, some images may appear excessively dark, providing little visibility. Consequently, images with abnormal data must be eliminated manually or automatically during the dataset production process. To address this, we apply an automatic filtering system to efficiently process and remove invalid images, thereby enhancing dataset reproducibility and facilitating collaboration among researchers. Abnormal images that require processing typically include those where severe environmental factors hinder the identification of people and equipment, images that capture only local features due to limited camera views or occlusion, repeated images, and severely dark images collected after mining operations have ceased. Additionally, blurred images from fast-moving targets during video-to-image conversion and images where distant targets are indistinguishable from other equipment due to environmental conditions and distance from the camera also fall into this category.

C. Data annotation

Finally, to label the filtered images, we employed the pretrained segment anything model (SAM) for automatic annotation tasks [23]. SAM is capable of generating high-quality object masks from input prompts such as points or boxes, making it versatile for various segmentation applications. Initially, SAM was fine-tuned on a limited number of manually annotated datasets of 510 images. Manual annotation was done using VGG image annotator[24]. Once fine-tuned, it was applied for automatic annotation across the remainder of the dataset. Additionally, we selected brighter images to create a subset for training an image enhancement component. To simulate low-light conditions, we introduced noise to this brighter subset by applying Gaussian noise, reducing brightness, and increasing contrast, thus generating low-light pairs of those images.

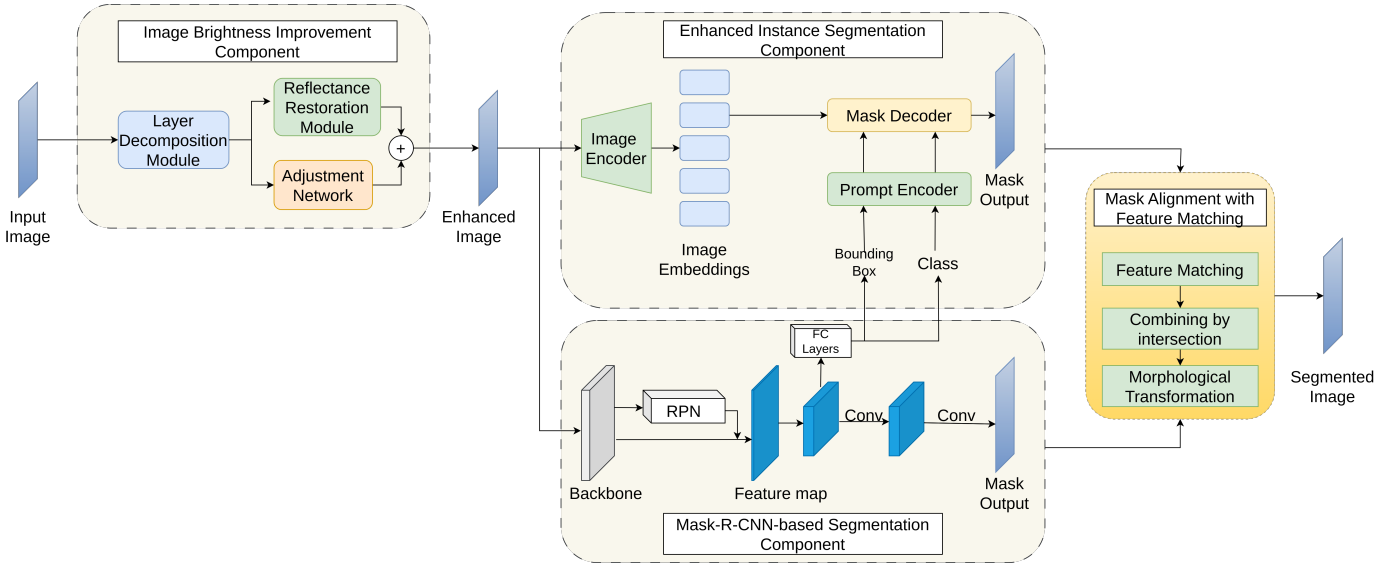


Fig. 3: An overview of proposed DIS-Mine instance segmentation framework.

IV. Proposed Methodology

This section provides an overview of the DIS-Mine framework (Fig. 3), which comprises three core components: Image Enhancement, Instance Segmentation (via the Segment Anything Model - SAM), and Mask R-CNN . The process flow of DIS-Mine is as follows:

- Each input image is processed through our proposed image brightness improvement network (Section IV-A), which is based on the KinD (Kindling the Darkness) network [12]. This network improves image quality by enhancing brightness, contrast, and texture. It works by decomposing, adjusting, and reconstructing the image for optimal enhancement.
- The improved image is used as input for the instance segmentation with SAM integration component (Section IV-B), which incorporates SAM as its core. The image is fed into the image encoder, while the bounding box and class prediction are passed into the prompt encoder to generate the instance mask.
- Concurrently, the improved image is also fed into our Mask R-CNN-based segmentation component (Section IV-C), which extends Faster R-CNN to generate bounding boxes, class predictions, and instance masks for each detected object.
- Finally, the outputs from both the instance segmentation with SAM integration component and the Mask R-CNN-based segmentation component are aligned through a series of operations, including feature matching, intersection-based combination, and morphological transformations, which are explained in detail in Section IV-D.

A. Image Brightness Improvement Component

The image brightness improvement component is developed to enhance the lighting of the ImageMine dataset, aiming for better prediction accuracy. Specifically, we integrate the KinD network [12] to the DIS-Mine network, which effectively enhances low-light images by improving visibility and quality in poorly lit environments. Grounded in Retinex theory [25], the KinD network decomposes images into reflectance and illumination components. The integrated network consists of three main modules: *the layer decomposition module* that separates the input image into reflectance and illumination layers; *the reflectance restoration module* that enhances the reflectance layer to recover true colors and details while minimizing noise; and *the adjustment network* that modifies the illumination layer to improve overall brightness and contrast. This image enhancement technique is applied to the input images before feeding them into the SAM model or Mask R-CNN model for training.

To begin with, we train this network on a synthetic normal/low-light dataset derived from relatively brighter images in our ImageMine dataset as we discussed in Section III-C. The network architecture includes three functional modules: i) layer decomposition, ii) reflectance restoration, and iii) illumination adjustment net. The normal/low-light pair images first pass through the layer decomposition block, which has two paths: one creates reflectance maps, while the other generates illumination maps. The reflectance path consists of five convolutional layers followed by a sigmoid layer. The first two convolutional layers are for downsampling, the next two are for upsampling, and the final convolutional layer is processed through the sigmoid layer to produce the reflectance map. In contrast, the illumination path consists of three convolutional layers and a sigmoid layer, utilizing features from the reflectance maps.

Following layer decomposition, the reflectance maps proceed to the reflectance restoration module, which includes a 5-layer U-Net [8], a convolutional layer, and a sigmoid layer. Simultaneously, the illumination maps go through an illumination adjustment layer, comprising two convolutional and rectified linear unit (ReLU) layers, and a convolutional

Algorithm 1: Training Process of Image Brightness Component

Input: Low-Light Image: I_{low} , Normal Image: I_{normal}
Output: Enhanced Image: I_{EN}

$RM_{\text{low}}, RM_{\text{normal}}, IM_{\text{low}}, IM_{\text{normal}} \leftarrow \text{LDM}(I_{\text{low}}, I_{\text{normal}})$
▷ `Extract Reflection Map (RM) and Illumination Map (IM) for low and normal images`
▷ `LDM indicates the layer decomposition module`
Calculate loss between $RM_{\text{low}}, RM_{\text{normal}}$
Calculate loss between $IM_{\text{low}}, IM_{\text{normal}}$
$RM_{\text{restored}} \leftarrow$ Reflectance Restoration $(RM_{\text{low}}, IM_{\text{low}})$
▷ `Restore reflection map based on low-light image`
$IM_{\text{adjusted}} \leftarrow$ Illumination Adjustment(IM_{low})
▷ `Adjust illumination map for better contrast`
Calculate loss between $RM_{\text{restored}}, RM_{\text{normal}}$
Calculate loss between $IM_{\text{adjusted}}, IM_{\text{normal}}$
Minimize the loss and update weights
$I_{\text{EN}} \leftarrow \text{Combine}(RM_{\text{restored}}, IM_{\text{adjusted}})$
Return I_{EN}

layer followed by a sigmoid layer. The outputs from both the reflectance restoration and illumination adjustment modules are combined pixel-wise to produce the final enhanced image, which increases light in relatively dark regions while maintaining brightness in lighter areas.

Algorithm 1 offers a explanation of the entire process.

B. Instance Segmentation with SAM Integration Component

We propose an enhanced version of the segment anything model (SAM) [23]. SAM is a promptable segmentation system that enables users to segment objects within an image using minimal input, such as a single click. It is designed to generalize to unfamiliar objects and images without the need for additional training. Specifically, SAM comprises three primary encoders: the image encoder, the prompt encoder, and the mask decoder.

The image encoder processes the input image to create an embedding that extracts relevant features, specifically adapted for high-resolution images. The prompt encoder accepts various types of prompts, distinguishing between sparse prompts (points, boxes, and texts) and dense prompts (masks). Finally, the mask decoder generates the segmentation mask based on the encoded image and prompts, processing the information through a down-sampling convolutional layer and concatenating it with the image embeddings. SAM has been trained on a massive dataset of 11 million images and 1.1 billion masks, enabling it to achieve strong zero-shot performance across a wide range of segmentation tasks. For our enhanced SAM version, we utilize only sparse prompts, which consist of boundary boxes and class labels predicted by the Mask R-CNN model (see Fig. 4). In Algorithm 2, we outline the process of our instance segmentation component.

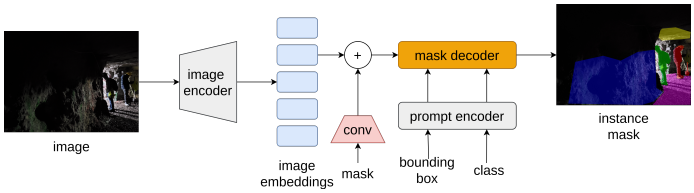


Fig. 4: Illustration of the instance segmentation with SAM integration component.

Algorithm 2: Training process of Instance Segmentation

Input: Enhanced Image: I_{EN}, Class Prediction: $Class$, Bounding Box: BB
Output: $Mask$

Image input for SAM, $I_{\text{i}} \leftarrow I_{\text{EN}}$
Prompt input for SAM, $P_{\text{i}} \leftarrow Class, BB$
$Mask \leftarrow \text{SAM}(I_{\text{i}}, P_{\text{i}})$
Return $Mask$

C. Mask R-CNN-based Segmentation Component

In this component, we extend the functionality of Mask R-CNN [14], which builds upon Faster R-CNN [15] by introducing an additional branch for predicting segmentation masks on each region of interest (RoI). This branch works alongside the existing branches for classification and bounding box regression, allowing the model to not only detect objects but also generate high-quality pixel-level masks for each detected instance.

The Mask R-CNN framework features two key innovations: RoIAlign and the mask branch. RoIAlign improves upon the RoIPool method used in Faster R-CNN by avoiding the quantization of RoI boundaries, ensuring precise spatial alignment, and preserving finer details. The mask branch is composed of a small fully convolutional network (FCN) that generates a binary mask for each RoI. Crucially, this mask branch operates independently from the class prediction, enabling more accurate and refined segmentation.

The overall loss of the Mask R-CNN model is divided into three parts. The first two losses represent the Fast R-CNN components: the classification loss and the bounding box regression loss, which can be expressed as follows:

$$\mathcal{L}_{\text{Fast R-CNN}} = \ell_{\text{class}} + \ell_{\text{box}} \tag{1}$$

where ℓ_{class} is the classification loss, which computes the difference between the predicted class probabilities and the ground truth labels using a softmax cross-entropy loss. ℓ_{box} is the bounding box regression loss, responsible for minimizing the difference between the predicted bounding box coordinates and the ground truth, typically using a smooth ℓ_1 loss. We can express (1) in a more generalized form as follows:

$$\ell(p_i, t_i) = \frac{1}{N_{\text{class}}} \sum_i \ell_{\text{class}}(p_i, p_i^*) + \lambda \frac{1}{N_{\text{box}}} \sum_i p_i^* \ell_{\text{box}}(t_i, t_i^*) \tag{2}$$

where N_{class} and N_{box} represent the number of ROIs used for classification and bounding box regression, respectively. The terms p_i and p_i^* correspond to the predicted and ground truth labels for the i-th ROI in the classification task, while

Algorithm 3: Training process of Mask R-CNN Component

Input: Enhanced Image: I_{EN}
Output: Instance Mask: $Mask$, Class Prediction: $Class$, Bounding Box: BB

Initialize Mask R-CNN model
Add modified mask loss to the multitask loss
Minimize the loss and update weights
$Mask, Class, BB \leftarrow$ Mask R-CNN(I_{EN})
Return $Mask, Class, BB$

t_i and t_i^* denote the predicted and ground truth bounding box parameters for the i-th ROI.

The third loss term, ℓ_{mask}, is unique to Mask R-CNN and enables the model to perform accurate pixel-wise segmentation in addition to object detection. This term allows Mask R-CNN to extend the capabilities of Faster R-CNN by not only identifying and locating objects but also generating precise segmentation masks for each instance. By introducing this mask loss, Mask R-CNN becomes a powerful tool, for instance, segmentation, transforming the traditional object detection framework into one capable of detailed pixel-level object delineation.

In our enhanced version of Mask R-CNN, we replace the standard binary cross-entropy mask loss ℓ_{mask} with a combination of weighted dice loss and focal loss. This modification is aimed at improving segmentation precision, particularly in cases of class imbalance or complex object boundaries. The new mask loss, combining these two functions, further enhances the model's instance segmentation performance. The modified loss can be expressed as:

$$\ell_{\text{mask}} = \ell_{w\text{-Dice}} + \ell_{\text{Focal}} \tag{3}$$

The Dice loss optimizes mask quality by ensuring overlap between predicted and ground truth masks. However, in cases of class imbalance, the weighted Dice loss, $\ell_{w\text{-Dice}}$, further improves segmentation by assigning a higher weight to the foreground class. This modification ensures that smaller or less frequent objects are segmented with greater accuracy. The weighted Dice loss, $\ell_{w\text{-Dice}}$, can be expressed as:

$$\ell_{\text{W-Dice}} = 1 - \frac{\sum_c w_c \cdot (2 \cdot \text{TP}_c)}{\sum_c (w_c \cdot (\text{TP}_c + \text{FP}_c + \text{FN}_c))} \tag{4}$$

where w_c is the weight for a class c, TP_c is the true positive count for class c, FP_c is the false positive count for class c, and FN_c is the false negative count for class c.

The focal loss ℓ_{focal}, reduces the influence of easy-to-learn examples, shifting focus towards challenging pixels, which are often prevalent in low-light or complex environments. This is particularly useful for improving segmentation performance in darker images similar to our dataset, ImageMine, where difficult examples dominate. The focal loss ℓ_{Focal} for binary classification is defined as:

$$\ell_{\text{Focal}} = -\alpha \cdot (1-\hat{p})^{\gamma} \cdot \log(\hat{p}) - (1-\alpha) \cdot \hat{p}^{\gamma} \cdot \log(1-\hat{p}) \tag{5}$$

where $\hat{p}$ is the predicted probability for the positive class, α is a weighting factor for the class, γ is the focusing parameter.

Algorithm 4: Mask Alignment Process

Input: Enhanced Image: I_{EN}
Output: Refined-Mask: $Mask_{\text{final}}$

$Mask_2, Class, BB \leftarrow$ Mask R-CNN-based Segmentation Component(I_{EN})
$Mask_1 \leftarrow$ Instance Segmentation component($I_i, Class, BB$)
Alignment using ORB feature-matching algorithm
$aligned\_mask_1 \leftarrow$ ORB_alignment($Mask_1, Mask_2$)
$aligned\_mask_2 \leftarrow$ ORB_alignment($Mask_2, Mask_1$)
$combined\_mask \leftarrow$ intersection($aligned\_mask_1, aligned\_mask_2$)
$dilated\_mask \leftarrow$ dilation($combined\_mask$)
$Mask_{\text{final}} \leftarrow$ erosion($dilated\_mask$)
Return $Mask_{\text{final}}$

Thus, the final loss function utilized by our enhanced Mask R-CNN is a multi-task objective that integrates three components: classification loss, bounding box regression loss, and mask prediction loss. This comprehensive loss function can be expressed as:

$$\mathcal{L}_{\text{total}} = \underbrace{\ell_{\text{class}} + \ell_{\text{box}}}_{\ell_{\text{Fast-R-CNN}}} + \underbrace{\ell_{w\text{-Dice}} + \ell_{\text{focal}}}_{\ell_{\text{enhanced-mask}}} \tag{6}$$

A detailed explanation of the entire process has been showed in Algorithm 3.

D. Mask Alignment with Feature Matching

This component is designed to select the most effective mask outputs from both segmentation models. Initially, the masks generated by the instance segmentation with SAM integration component and the Mask R-CNN-based segmentation model undergo an alignment process using a feature-matching algorithm, specifically the oriented fAST and rotated brief (ORB) method [26]. After alignment, we combine the masks by taking their intersection, highlighting the common areas identified by both models.

Next, we refine the final instance mask output using morphological operations that combine dilation and erosion. Additionally, we implement rule-based operations based on the mine's structure. We divide the image into 4x4 grids and classify instances according to their positions within these grids. For example, any object detected in the lower row of the grid is designated as part of the road class, even if it extends into the upper grid. The complete process of this component is detailed in Algorithm 4.

V. Experiments and Results

A. Experimental setup

Datasets. To evaluate DIS-Mine against state-of-the-art approaches, we consider a couple of benchmark datasets for poor-light conditions segmentation tasks, including the LIS dataset [10], DsLMF+ [27], along with our collected dataset, ImageMine (for more information about this dataset, please refer to Section III). Below is a brief description of other datasets:

TABLE I: Simulation Parameters

Hyperparameter	DIS-Mine Components		
	Image Improvement	Optimized SAM	Mask R-CNN-based
Learning Rate	0.001	0.0001	0.001
Batch Size	16	32	4
Optimizer	Adam	Adam	SGD
Epochs	100	80	80
Loss Function	MSE	Cross-Entropy	Multi-task Loss
Backbone	-	ViT	ResNet101
Image Size	256x256	1024x1024	1024x1024
Augmentation	Random Crop, Flip	Random Crop	Random Crop
Regularization	L2 Regularization	Dropout	L2 Regularization

- **LIS dataset[10].** This dataset comprises 2,230 pairs of low-light and normal-light images collected from diverse indoor and outdoor scenes. It aims to tackle the challenges of instance segmentation in extremely low-light conditions, where traditional methods often struggle. In our experiments, we specifically focused on the compelling low-light images from this dataset.
- **DsLMF+[27].** This dataset consists of 138,004 images captured in underground longwall mining faces, encompassing six classes: mine personnel, hydraulic support guard plates, large coal, towlines, miners' behaviors, and mine safety helmets.

Training Procedure. The training procedure of DIS-Mine encompasses the training of each individual component. The hyperparameters used for training each component are summarized in Table I. For the automatic annotation task, we employed the optimized SAM model, training it with pre-trained model weights on manually annotated images. Masks for each image were generated using the optimized SAM model, and these image-mask pairs were subsequently utilized to train our third component, the Mask R-CNN model. Additionally, the SAM model was retrained on both manually and automatically annotated datasets.

Environment. We conducted our training on a DELL R740xa equipped with 238 GB of RAM, an Nvidia A100 GPU, and 80 GB of GPU memory. Various versions of Python 3 and packages from the PyTorch framework were utilized.

B. Baselines

We compare DIS-Mine with several state-of-the-art instance segmentation methods, including SAM [23], Mask R-CNN [14], Mask2Former [28], and Instance Segmentation in Dark (ISD) [10]. These models serve as strong baselines in the field of semantic segmentation, particularly under poor lighting conditions, allowing for a comprehensive assessment of the effectiveness of DIS-Mine. Below is a brief description of each model:

- **SAM [23].** SAM is a versatile prompt-based segmentation model capable of generating instance masks across various contexts.
- **Mask R-CNN [14].** Mask R-CNN is known for its robust performance in object detection and instance segmentation, extending Faster R-CNN with a pixel-level mask prediction branch.
- **Mask2Former [28].** Mask2Former is a universal segmentation model for instance, semantic, and panoptic tasks, using masked attention to improve segmentation accuracy.
- **ISD [10].** ISD is designed for low-light conditions, employing adaptive downsampling and disturbance-suppressing learning to reduce noise.

Evaluation metrics. We assess the performance of DIS-Mine against our baselines using two widely adopted evaluation metrics in instance segmentation: F1-score and mean intersection over union (mIoU). We choose these metrics due to their effectiveness in measuring both segmentation accuracy and overlap between predicted and ground truth masks. Below is a brief description of each metric:

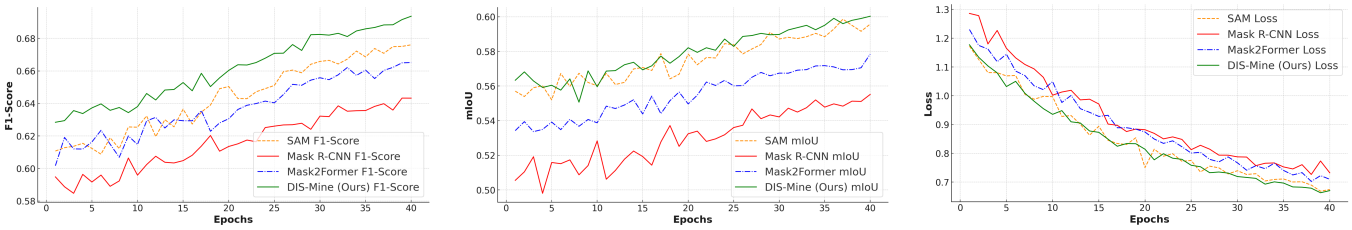
- **F1-Score:** The F1-Score is the harmonic mean of precision and recall, providing a balanced metric that accounts for both false positives and false negatives. In instance segmentation, precision measures how accurately the predicted instance masks align with ground truth masks, while recall assesses how effectively the model detects all relevant instances. A high F1-Score indicates that the model achieves both high precision (few false positives) and high recall (few false negatives), reflecting strong overall performance in detecting and segmenting objects. The mathematical formula for the F1-Score is:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

- **IoU & mIoU.** is a measure of overlap between the predicted mask and the ground truth mask. It is defined as the area of overlap between the predicted and actual instance masks divided by the area of their union. IoU provides a direct assessment of how accurately the model predicts the shape and boundaries of instances. A higher IoU score reflects better segmentation quality, as it indicates more accurate pixel-level alignment between prediction and ground truth. mIoU (mean IoU) refers to the average of IoU values over all classes in a dataset. It measures the overlap between the predicted segmented mask and ground truth for each class which summarizes the model's segmentation accuracy. The mathematical formula for the mIoU is:

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \tag{8}$$

where N is the total number of classes, A_i and B_i represent the predicted and ground truth areas of class i. $|A_i \cap B_i|$ is the area of overlap between the predicted segmentation A_i and the ground truth B_i for class i. $|A_i \cup B_i|$ is the total area covered by both the predicted segmentation and the ground truth for class i.



(a) F1-score. (b) mIoU score. (c) Loss-Curve.

Fig. 5: Evaluation of DIS-Mine against baselines on various metrics on the ImageMine dataset.

C. Results

Comparison with SOTA. Table II presents a comparison of our proposed approach, DIS-Mine, against baseline models across three different datasets, including our own, ImageMine. It is important to note that the ISD model was only evaluated on the LIS-dataset, as it requires clean, high-quality images for training, which limits its application to other datasets. The results provide a detailed analysis of two evaluation metrics: F1-score and mIoU. On the ImageMine dataset, DIS-Mine achieves the highest performance with a 70.2% F1-score and 60.5% mIoU, followed by the SAM model, which obtains 68.7% F1-score and 60.0% mIoU. For the LIS dataset, while ISD excels in low-light conditions due to its specialized training, DIS-Mine again outperforms all baselines it in the F1-score, demonstrating its adaptability to challenging environments, with ISD emerging as the second-best model. On the DsLMF+ dataset, DIS-Mine achieves the top performance with an F1-score of 86% and mIoU of 72%, with SAM again ranking second. These results demonstrate the robustness of DIS-Mine by achieving a strong balance between F1-score and mIoU across multiple datasets, highlighting its generalizability for diverse instance segmentation tasks.

Furthermore, in Fig. 5, we show the performance trends of our DIS-Mine approach against the baselines over the training period on our ImageMine dataset, evaluating metrics like F1-Score, mIoU, and Loss. DIS-Mine's F1-Score curve, which plateaus at a higher value, suggests more accurate and consistent segmentation predictions. Similarly, our DIS-Mine achieves higher mIoU demonstrating its strong alignment between predictions and actual instances. The loss curves, on the other hand, reveal the models' optimization processes, where a downward trend represents error reduction and convergence toward more accurate predictions. DIS-Mine's lower final loss value reflects better optimization, with reduced prediction errors compared to the other models.

Among the other baselines, Mask2Former consistently ranks as the second-best model in terms of mIoU and F1-Score after our DIS-Mine, indicating good segmentation accuracy but with somewhat slower convergence and less effective boundary handling compared to DIS-Mine. SAM and Mask R-CNN show relatively weaker performance, with SAM achieving moderate segmentation accuracy and Mask R-CNN displaying higher loss values and slower convergence. These limitations may stem from their challenges in optimizing boundary alignment and balancing precision with recall.

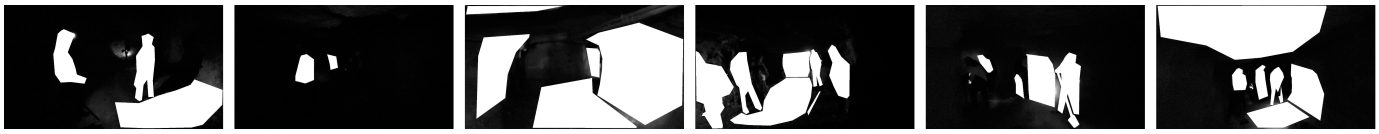
TABLE II: Comparing DIS-Mine vs. baselines across various datasets including our dataset, ImageMine.

Dataset	Model	F1-score	mIoU
ImageMine (Ours)	SAM	68.7%	60.0%
	Mask R-CNN	65.0%	56.0%
	Mask2Former	67.2%	58.0%
	DIS-Mine (ours)	70.2%	60.5%
LIS-dataset	SAM	61.0%	47.5%
	Mask R-CNN	58.0%	44.6%
	Mask2Former	62.0%	45.8%
	ISD	61.7%	49.8%
	DIS-Mine (ours)	63.2%	47.0%
DsLMF+	SAM	84.0%	71.0%
	Mask R-CNN	80.0%	68.0%
	Mask2Former	83.0%	72.0%
	DIS-Mine (ours)	86.0%	72.0%

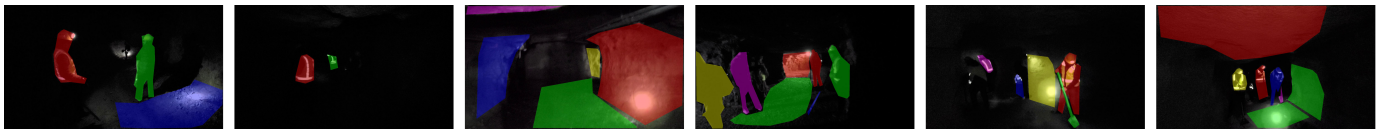
Analyzing each curve's slope, plateau, and final values offers insight into the models' learning dynamics. DIS-Mine stands out with its faster convergence, balanced precision-recall performance, and superior segmentation accuracy. Overall, these performance curves highlight DIS-Mine's clear advantage in instance segmentation, demonstrating its effectiveness in both accuracy and optimization relative to the other models.

Class-Wise Performance Analysis. In this study, we evaluate the performance of the DIS-Mine by focusing on six object classes: people, equipment, corridor, road, wall, and roof. The study involves systematically removing one or more of these classes during training to assess the impact on overall segmentation accuracy. By excluding specific classes in different experiments, we aim to analyze the model's reliance on particular object categories and how their inclusion or exclusion affects the segmentation accuracy of the remaining classes. For example, the removal of dynamic classes such as people, equipment, and corridors enables us to examine the model's performance on static, large-scale structures like

(a) Samples from our ImageMine dataset (input images).

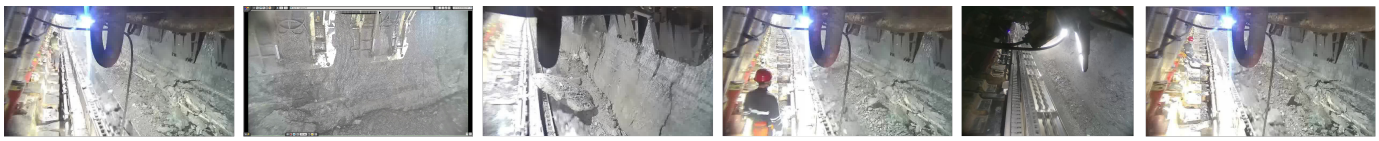


(b) Corresponding ground-truth masks.



(c) Generated predicted masks.

Fig. 6: DIS-Mine prediction results on samples from our collected ImageMine dataset.



(a) Samples from DsLMF+ dataset (input images).

(b) Generated predicted masks.

Fig. 7: DIS-Mine prediction results on samples from DsLMF+ dataset.

roads, walls and roofs, thereby assessing its ability to generalize without the presence of more complex or variable object classes. This methodology provides insight into the interdependence between object classes and highlights the role of class diversity in promoting the robustness and generalization capacity of the model. Our findings indicate that the model faces significant challenges when segmenting geometrically similar classes, such as roads, walls, and roofs. These classes share similar color and contrast characteristics, and due to the lack of distinctive edges, they often become indistinguishable in certain cases. This highlights the model's limitations in differentiating between visually analogous categories. As a result, we merged the road, roof, and wall classes into a single class called *surrounding*. Table III presents the performance of each class based on this merging approach.

Finally, in Fig. 6 and Fig. 7, we showcase the instance segmentation outputs generated by our DIS-Mine approach on samples from the ImageMine dataset and the DsLMF+ dataset. The results highlight DIS-Mine's effectiveness in accurately segmenting instances under challenging low-light conditions. Specifically, the results in Fig. 6 demonstrate that DIS-Mine's predictions closely align with the ground truth masks for various scenarios involving people, equipment, and corridors, yielding meticulous and detailed segmentation outputs. Similarly, the results in Fig. 7 show that DIS-Mine provides reliable predictions for various classes in the DsLMF+ dataset, including coal miners, hydraulic supports, large coal, miner safety helmets, towlines, and miners' behaviors.

TABLE III: Performance based on individual class for DIS-Mine on ImageMine

Class	F1-score	IoU
People	72.6%	72.6%
Equipment	71.4%	62.1%
Corridor	88.3%	78.5%
Surrounding	64.0%	52.4%

VI. Conclusion and Future Work

In this paper, we proposed a novel instance segmentation method named DIS-Mine to segment images of underground

mines in poor-light conditions. By enhancing the images using KinD, utilizing specific prompts generated from Mask R-CNN, and aligning the masks from both the optimized SAM and Mask R-CNN models, our approach leverages the strengths of both models to achieve more accurate and robust segmentation in poor lighting conditions. The results highlight the potential of model integration and mask alignment in overcoming the limitations posed by poor light conditions because of high noise, and poor contrast. Most importantly, we collected a real-world underground mine dataset in very dark conditions called ImageMine to validate our results and compare our method with other models' performance. Our comprehensive results show that DIS-Mine outperforms each baseline model individually, delivering significant improvements in both F1-score and mIoU. Additionally, we evaluated DIS-Mine's performance on multiple low-light datasets to demonstrate the generalizability of our method across various challenging segmentation tasks.

For future work, we aim to further enhance our approach by incorporating multimodal data, such as thermal imaging and LiDAR point cloud data, to generate even more precise segmentations in complex scenarios. This integration of multimodality has the potential to significantly improve segmentation accuracy by providing complementary information, particularly in extreme low-light or adverse environmental conditions. By utilizing these diverse data sources, we hope to develop a more robust framework capable of performing in a wider range of real-world applications.

ACKNOWLEDGEMENT

We want to extend our gratitude to the members of the W2C lab who have participated in the data collection in the underground mine. This research is supported by a grant from CDC-NIOSH.

REFERENCES

[1] L. H. C. D. M. Debing, "Current status of deep mining and disaster prevention in china," *Coal Science and Technology*, no. 1, 2016.

[2] Y. Zhang *et al.*, "Residual coal exploitation and its impact on sustainable development of the coal industry in china," *Energy Policy*, vol. 96, pp. 534–541, 2016.

[3] W. P. Rogers *et al.*, "Automation in the mining industry: Review of technology, systems, human factors, and political risk," *Mining, metallurgy & exploration*, vol. 36, pp. 607–631, 2019.

[4] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta, "Tweedr: Mining twitter to inform disaster response." in *ISCRAM*, 2014, pp. 269–272.

[5] S. Karlsson, B.-I. Saveman, M. Hultin, U. Björnstig, and L. Gyllencreutz, "Preparedness for peer first response to mining emergencies resulting in injuries: a cross-sectional study," *BMJ open*, vol. 10, no. 11, 2020.

[6] M. Onifade, "Towards an emergency preparedness for self-rescue from underground coal mines," *Process Safety and Environmental Protection*, vol. 149, pp. 946–957, 2021.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[8] O. Ronneberger, P. Fischer, and T. 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*. Springer, 2015, pp. 234–241.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[10] L. Chen, Y. Fu, K. Wei, D. Zheng, and F. Heide, "Instance segmentation in the dark," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 2198–2218, 2023.

[11] W. Yang *et al.*, "Advancing image understanding in poor visibility environments: A collective benchmark study," *IEEE Transactions on Image Processing*, vol. 29, pp. 5737–5752, 2020.

[12] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1632–1640.

[13] A. Arnab and P. H. S. Torr, "Pixelwise instance segmentation with a dynamically instantiated network," 2017.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[16] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Hybrid task cascade for instance segmentation," 2019.

[17] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9157–9166.

[18] E. Mohamed, A. Shaker, A. El-Sallab, and M. Hadhoud, "Insta-yolo: Real-time instance segmentation," 2021.

[19] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015.

[20] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[21] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *CoRR*, vol. abs/2005.12872, 2020.

[22] J. Lin, N. Anantrasirichai, and D. R. Bull, "Feature denoising for low-light instance segmentation using weighted non-local blocks," *ArXiv*, vol. abs/2402.18307, 2024.

[23] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[24] A. Dutta, A. Gupta, and A. Zissermann, "VGG image annotator (VIA)," http://www.robots.ox.ac.uk/ vgg/software/via/, 2016.

[25] J. McCann, *Retinex Theory*. New York, NY: Springer New York, 2016, pp. 1118–1125.

[26] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.

[27] X. Zhang, W. Yang, B. Ma, and Y. Wang, "Dslmf+: An open dataset for intelligent recognition of abnormal condition in underground longwall mining face," 2024.

[28] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," *arXiv*, 2021.