ISPRS Journal of Photogrammetry and Remote Sensing 183 (2022) 470–481

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

# LISU: Low-light indoor scene understanding with joint learning of reflectance restoration

Ning Zhang *, Francesco Nex, Norman Kerle, George Vosselman

*Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, the Netherlands*

ARTICLE INFO

*Keywords:*
Semantic segmentation
Deep learning
Intrinsic image decomposition
Low-light

ABSTRACT

Semantic segmentation using convolutional neural networks (CNNs) achieves higher accuracy than traditional methods, but it fails to yield satisfactory results under illumination variants when the training set is limited. In this paper we present a new data set containing both real and rendered images and a novel cascade network to study semantic segmentation in low-light indoor environments. Specifically, the network decomposes a low-light image into illumination and reflectance components, and then a multi-tasking learning scheme is built. One branch learns to reduce noise and restore information on the reflectance (reflectance restoration branch). Another branch learns to segment the reflectance map (semantic segmentation branch). The CNN features from two tasks are concatenated together so as to improve the segmentation accuracy by embedding the illumination-invariant features. We compare our approach with other CNN-based segmentation frameworks, including the state-of-the-art DeepLab v3+, on the proposed real data set, and our approach achieves the highest mIoU (47.6%). The experimental results also show that the semantic information supports the restoration of a sharper reflectance map, thus further improving the segmentation. Besides, we pre-train a model with the proposed large-scale rendered images and then fine-tune it on the real images. The pre-training results in an improvement of mIoU by 7.2%. Our models and data set are publicly available for research. This research is part of the EU project INGENIOUS[1]. Our data sets and models are available on our website[2].

## 1. Introduction

Semantic segmentation is a basic task of computer vision and aims at classifying each pixel of an image. Compared with the original RGB image a segmentation map presents the scene in a more intuitive way, and it is useful in many fields, such as remote sensing (Kemker et al., 2018), autonomous driving (Neuhold et al., 2017), and vision-based indoor navigation of robots or unmanned aerial vehicles (UAVs) (Adachi et al., 2019; Gupta et al., 2020; Lu et al., 2018). In the EU project INGENIOUS we aim at increasing first responders' situational awareness in rescue operations. Specifically, first responders need to look inside buildings to search for trapped people or verify the presence of possible dangers. Therefore, we focus on developing semantic segmentation algorithms integrated with UAVs to be used in indoor environments. Thanks to many open source annotated data sets such as NYU-depth v2 (Silberman et al., 2012), SUN RGBD (Song et al., 2015), and SceneNet (Handa et al., 2016), semantic segmentation using convolutional neural networks (CNNs) has dominated and achieved better performance than traditional methods (Garcia-Garcia et al., 2017). However, CNN-based models are limited by the distribution of training sets, hence they are not robust to illumination changes. When indoor robots or small UAVs explore the rooms without electricity after a disaster, they usually illuminate the scene with LED lights (Özaslan et al., 2015; Lau and Ko, 2007). Therefore, an object will display different colors and textures when camera views change, and there may be shadows on the surface of the object, which makes the network trained on a limited training set unable to predict correct semantic labels. It is important to study how to make CNNs more robust to illumination changes when it is unrealistic to obtain a complete enough data set for training. Some researchers used preprocessing steps on the original images and obtain illumination-invariant images to overcome the adverse impact of illumination changes (Alshammari et al., 2018; Maddern et al., 2014; Upcroft et al., 2014). Nevertheless, these methods sometimes fail because they are sensitive to the saturation of images (Upcroft et al., 2014).

* Corresponding author.
*E-mail addresses:* n.zhang@utwente.nl (N. Zhang), f.nex@utwente.nl (F. Nex), n.kerle@utwente.nl (N. Kerle), george.vosselman@utwente.nl (G. Vosselman).

[1] https://ingenious-first-responders.eu/

[2] https://github.com/noahzn/LISU

https://doi.org/10.1016/j.isprsjprs.2021.11.010
Received 20 April 2021; Received in revised form 13 October 2021; Accepted 17 November 2021

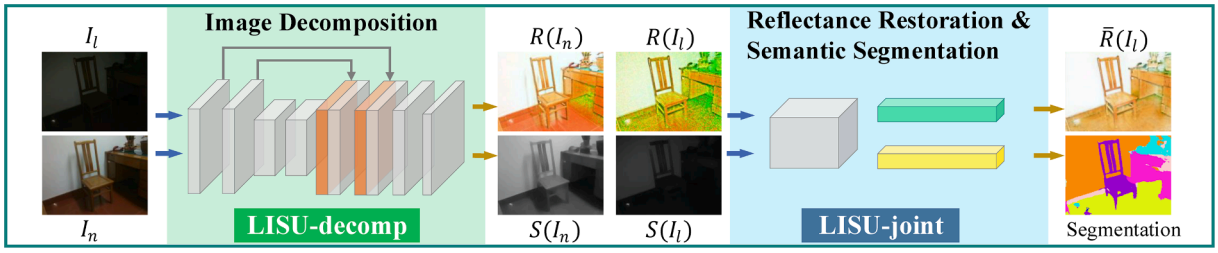


**Fig. 1.** The cascade architecture of our LISU network, which consists of LISU-decomp and LISU-joint. For training the input is a pair of low-light and normal-light images. LISU-decomp learns the decomposition of each input image, and LISU-joint jointly learns the reflectance restoration and semantic segmentation. Each cube represents the feature map generated by a convolution layer.

This paper attempts to simulate a UAV with LED to explore a dark room. We focus on semantic segmentation in real low-light indoor scenes, and use illumination-invariant CNN features to improve the accuracy of segmentation. Retinex theory proposed by Land and McCann (1971) studied color constancy and was further developed to solve the intrinsic image decomposition problem. An image can be decomposed as the product of an illumination component and a reflectance component. The reflectance component represents the intrinsic property of an object like true colors and textures and is unchanged with illumination variants. Therefore, the reflectance component of an image favors the semantic segmentation task. In this paper, we propose a novel cascade framework LISU, for **L**ow-light **I**ndoor **S**cene **U**nderstanding. As shown in Fig. 1, the proposed framework consists of two parts, one is an unsupervised decomposition network LISU-decomp that decomposes RGB images into corresponding illumination maps and coarse reflectance maps. And the other is an encoder-decoder network LISU-joint to learn reflectance restoration and semantic segmentation in a multi-tasking way. We fuse the feature maps from two tasks together for tighter joint learning. The contributions of this paper can be summarized as follows:

- We propose a novel cascade framework LISU to improve the segmentation accuracy of low-light indoor scenes by joint learning of semantic segmentation and reflectance restoration.
- To study the specific segmentation task under low-light, we present the first low-light indoor scene understanding data set. The data set consists of a large-scale realistic rendered data set and a small-scale real data set. In the data set pixel-wise annotations and depth maps are provided. We hope that this data set can make researchers pay more attention to the task of low-light indoor scene understanding.
- The experimental results show that both reflectance restoration and semantic segmentation tasks benefit from the joint learning. The illumination-invariant features help the segmentation task. Meanwhile, the semantic information also supports the restoration of a sharper reflectance map, thus further improving the segmentation. Our model and data set will be available online for research.

The remainder of the paper is organized as follows. Relevant recent research is summarized in Section 2. The structure of the proposed LISU framework is presented in Section 3. Section 4 introduces our data set in detail. Section 5 introduces the evaluation metrics and implementation details. The experimental results and discussion are elaborated in Section 6. We conclude this paper in Section 7.

## 2. Related work

This section reviews some recent related work, particularly the intrinsic image decomposition, indoor semantic segmentation and multi-task learning.

### 2.1. Intrinsic image decomposition

Intrinsic image decomposition reconstructs an image $I$ in the form of a product of a reflectance map $R(I)$ and an illumination map $S(I)$:

$$I = R(I) \cdot S(I). \tag{1}$$

A perfect decomposition enables us to obtain the reflection map containing the inherent colors and textures of the scene. However, the decomposition task is ill-posed because there are countless combinations of reflectance and illumination maps that can reconstruct the same image. To address this problem researchers have explored to impose additional constraints on decomposed components. The pioneering Retinex theory (Land and McCann, 1971) assumes that the illumination component is smooth and only yields small gradients. Some subsequent studies also proposed different priors to guide the decomposition. Rother et al. (2011) achieved competitive decomposition results by modelling the reflectance component using basic colors and integrating additional edge information. Shen et al. (2008) designed a texture prior to the reflection, so that the reflection values with the same textures are continuous. Tappen et al. (2005) trained a classifier to classify image derivatives into shading or reflectance according to color and gray-scale information. With the rise of convolutional neural networks (CNNs) in recent years, many supervised methods have been proposed to decompose images in synthetic data sets (Fan et al., 2018; Li and Snavely, 2018; Narihira et al., 2015). Since it is quite difficult to annotate real images for decomposition tasks, some researchers tried unsupervised learning. Janner *et al.* first decomposed an input image, and then used a trained shading model to reconstruct the image using decomposed components. Therefore, they can minimize an unsupervised reconstruction error to update the decomposition network (Janner et al., 2017). Liu et al. (2020) proposed to use physical and domain constraints to train an unsupervised decomposition network from uncorrelated data.

The decomposed reflectance and illumination can be further processed and used for other specific application such as low-light image enhancement. Guo et al. (2016) estimated and refined the illumination map by applying a structure prior. Wei et al. (2018) trained a CNN-based Retinex-Net to decompose low/normal-light image pairs into a shared reflectance map and enhance the illumination map of the low-light image. Similarly, (Zhang et al., 2019) also used image pairs for training, but additionally explored to use normal-light images to restore the reflectance maps of low-light images. Zhu et al. (2020) designed a novel loss function to remove noise from the reflectance map. What these low-light enhancement papers have in common is that they use the intrinsic image decomposition model to obtain intermediate decomposition of a low-light image, and then refine the degraded reflectance and adjust the illumination map. In this paper, our focus is not to enhance the low-light images, but to use the decomposed components for a joint learning of semantic segmentation and reflectance restoration.

### 2.2. Semantic segmentation

CNN-based semantic segmentation has achieved great success since Long et al. proposed a fully convolutional network (FCN) for pixel-wise classification (Long et al., 2015). Since then, most CNN-based semantic segmentation networks have adopted the basic idea of FCN, but put forward different strategies to improve the accuracy of segmentation. SegNet used an encoder-decoder structure and introduced pooling indices for nonlinear up-sampling (Badrinarayanan et al., 2017). U-Net (Ronneberger et al., 2015) links the shallow features from the encoder to corresponding decoder layers and achieved good results in medical image segmentation. Refine-Net presented by Lin et al. (2017) aggregated multi-level features and obtained great improvement in segmentation accuracy. Some work utilized depth maps as additional geometry information to augment the indoor semantic segmentation (Cheng et al., 2017a; Park et al., 2017). Cheng et al. (2017a) proposed a gated fusion layer to fuse the geometric information and refined boundary segmentation using a deconvolution structure. Park et al. (2017) introduced a novel fusion block to study an optimal fusion of RGB and depth features. Researchers have created many dedicated real data sets for indoor scene understanding, such as NYU-Dv2 (Silberman et al., 2012) and SUN RGBD (Song et al., 2015). However, collecting and labeling large-scale real images for training is laborious. Owing to the development of computer graphics, synthetic data sets have been created to pre-train CNNs (Handa et al., 2016; McCormac et al., 2017). To our knowledge, these data sets only focus on normal-light scenes.

### 2.3. Semantic segmentation of low-light images

This is a relatively new and challenging computer vision task, one that has gradually attracted public attention in vision-based autonomous driving at night. To address this problem some work adapted the model trained on off the shelf normal-light data sets to low-light scenes. Dai and Van Gool, (2018) proposed to leverage transfer learning to adapt a model trained on daytime scenes to nighttime scenes. Sakaridis et al. (2019) adapted the model progressively and they did not use any nighttime annotations. Sun et al. (2019) used a generative adversarial network (GAN) to convert nighttime images to daytime images, hence they could make use of the existing models trained on daytime images. This similar method was also used by Cho et al. (2020) who proposed a modified CycleGAN to translate low-light images for road scene segmentation. However, it is not easy to train an adversarial network if two domains have great gaps. Other work explored transforming RGB images into illumination-invariant images for more robust semantic segmentation (Alshammari et al., 2018; Maddern et al., 2014; Upcroft et al., 2014). Xu et al. (2019) designed a system to assist visually impaired people and they pre-processed the training set in a similar way as described above.

With the development of small indoor UAVs or robots, they have been used for indoor inspection (Giernacki et al., 2017; Li et al., 2018; Kwon et al., 2008). These devices are also helpful in post-disaster indoor rescue missions. They can explore low-light indoor environments without a power supply by using LED lights and cameras, and automatically navigate using semantic information of scenes if integrated with advanced scene understanding algorithms. However, aforementioned works focused on outdoor scene semantic segmentation for autonomous driving field, but there is little research on low-light indoor scene semantic segmentation. In this paper we try to fill the research gap of low-light indoor scene understanding.

### 2.4. Multi-task learning (MTL)

MTL learns two or more closely related tasks and optimizes the model together. Zhang and Yang (2021) pointed out that MTL helps knowledge flow in different task branches, so it can build a more robust deep network. Much CNN-based work benefited from utilizing relevant features from different tasks and achieved better results. Dai et al. (2016) proposed to simultaneously learn instance segmentation, mask segmentation, and bounding box regression. All three tasks are related to extracting semantic information, so the introduction of MTL can make the network have stronger semantic representation ability. Baslamisli et al. (2018) created a synthetic data set of daytime natural environments, and jointly trained a supervision network for semantic segmentation and intrinsic image decomposition. Jiao et al. (2019) used an encoder to learn the shared features of depth estimation and indoor semantic segmentation, and then utilized the distilled geometry information to guide the segmentation task. SegFlow (Cheng et al., 2017b) learned segmentation and optical flow jointly by propagating features of these two tasks between two branches, and achieved better results for both tasks. (Zeng et al., 2019) used a similar approach to jointly learn saliency detection and semantic segmentation. This paper adopts the idea of MTL and we restore the degraded reflectance maps and segment the scenes simultaneously.

## 3. LISU: a new framework for low-light indoor semantic segmentation

In this section the proposed LISU framework for low-light indoor scene understanding is illustrated in detail. As shown in Fig. 1 it is a cascade network and consists of two sub-networks. A low-light image is first fed into the first network LISU-decomp and decomposed into a reflectance map and an illumination map. Then, they are inserted into the second network, LISU-joint, for a multi-task learning, which is joint learning of semantic segmentation and reflectance restoration. In the training phase LISU-decomp takes paired low-/normal-light images as input and decomposes images in an unsupervised way, *i.e.,* it does not use the ground-truth of reflectance or illumination, but uses the inherent attributes of images for training. We introduce the details of the unsupervised training in Section 3.1. In the testing phase only low-light images are needed.

Baslamisli et al. (2018) proposed a fully supervised framework to learn intrinsic images and semantic segmentation simultaneously, and they used synthetic daytime outdoor garden images for training. However, we cannot use their method because we focus on inspection and rescue missions in real low-light indoor scenes (with LED lighting), and we do not have the ground-truth of reflectance and illumination of real images for training. Therefore, we design this novel cascade framework to deal with the real-world low-light indoor semantic segmentation task.

In the following sub-sections we present the detailed structures and loss functions of LISU-decomp and LISU-joint.

### 3.1. LISU-decomp: intrinsic image decomposition

Our decomposition network takes a three-channel RGB image as input, and outputs a three-channel reflectance map and a one-channel illumination map. It uses long skip connection to bridge the features from encoder layers to decoder layers. Since this connection strategy can preserve low-level information and yield sharper features, it has been used both in semantic segmentation (Ronneberger et al., 2015; Wu et al., 2018) and intrinsic image decomposition (Dai et al., 2016; Rematas et al., 2016).

Since we are not able to get the ground-truth of intrinsic image decomposition of real images, we follow the unsupervised decomposition approach proposed by Zhang et al. (2019) and we input paired low-light and normal-light images of a same scene $[I_l, I_n]$. LISU-decomp decomposes them into reflectance maps and illumination maps, namely $[R(I_l), S(I_l)]$ and $[R(I_n), S(I_n)]$, respectively. We define the first part of the reconstruction loss according to Eq. 1:

$$L_{recon1} = \| I_l - R(I_l) \cdot S(I_l) \|_1 + \| I_n - R(I_n) \cdot S(I_n) \|_1, \quad (2)$$

where $\| \cdot \|_1$ denotes the $l^1$ norm. According to the definition of reflec-


**Fig. 2.** Comparison of reflectance maps of low/normal-light images output by LISU-decomp. Degradation and noise are more obvious in the reflectance map of the low-light image.

tance, the reflectance maps of two images taken under different lighting conditions in the same scene should be equal. Thereby, an ideal decomposition network enable us to reconstruct the normal-light image using the illumination map of the normal-light image and the reflectance map of the low-light image, and vice versa. The second part of the reconstruction loss can be defined as follows:

$$L_{recon2} = \| I_l - R(I_n)\cdot S(I_l)\|_1 + \| I_n - R(I_l)\cdot S(I_n)\|_1. \quad (3)$$

Some Retinex-based decomposition methods extract the maximum value in three channels of RGB image as a preliminary illumination map (Guo et al., 2016; Handa et al., 2016; Land and McCann, 1971). We

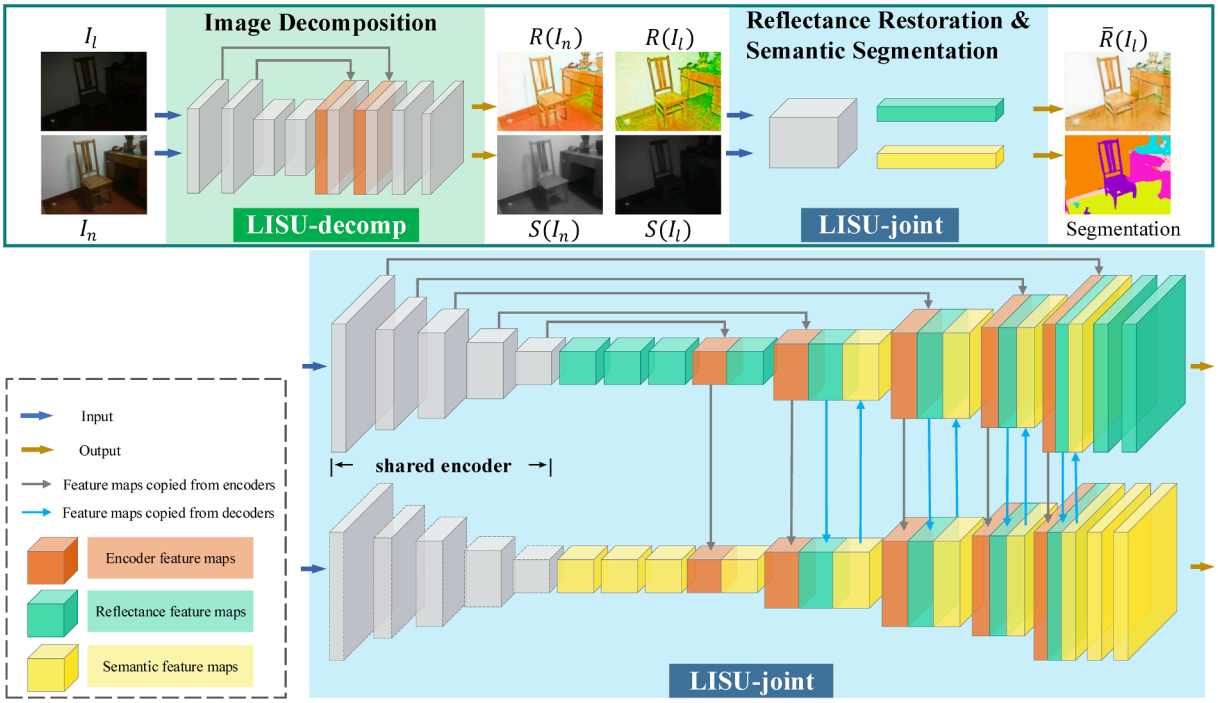


**Fig. 3.** The detailed structure of the joint learning network LISU-joint.


**Fig. 4.** Sample images from the proposed LLRGBD data sets. The images in the first two rows are synthetic, and the last two rows are real images.

follow this and define a loss function constraining the illumination map:

$$L_{init} = \| S\left(I_l\right) - \max_{c\in\{R,G,B\}} I_l^c \|_1 + \| S\left(I_n\right) - \max_{c\in\{R,G,B\}} I_n^c \|_1. \quad (4)$$

Inspired by Wei et al. (2018) we also define a structure-awareness smoothness loss on the illumination map:

$$L_{smooth} = \| \nabla S(I_l) \cdot exp\left(-\lambda_g \nabla I_l\right) \|_1 + \| \nabla S(I_n) \cdot exp\left(-\lambda_g \nabla I_n\right) \|_1, \quad (5)$$

where $\nabla$ computes the first-order derivative of horizontal and vertical directions. $\lambda_g$ is set to 10 as in Wei et al. (2018) to work as weighing the smooth continuity. Different from Wei et al. who used reflectance as a weight term, we try to find clues from the input low-light images because the reflectance map obtained at this stage is too noisy to guide a satisfactory decomposition. The combined loss function to train the decomposition network can be expressed as:

$$L_{decomp} = L_{recon1} + \lambda_1 L_{recon2} + \lambda_2 L_{init} + \lambda_3 L_{smooth}, \quad (6)$$

where $\lambda_1, \lambda_2$ and $\lambda_3$ are weight factors. We study and discuss different weights combinations in Section 6.4.

### 3.2. LISU-joint: joint learning of reflectance restoration and semantic segmentation

Since reflectance maps are invariant to illumination changes, the natural idea is to use the reflectance map obtained at the first stage as the input to the segmentation network. However, as shown in Fig. 2, the unsupervised decomposition of a low-light image generates very poor quality reflectance maps and a lot of information is lost. Inspired by Zhang et al. (2019) who restored the degraded reflectance map with the reflectance map of the normal-light image, we extend the single segmentation network to a multi-task network by adding another branch to restore the reflectance. Although the reflectance map of a normal-light image also has noise, we use it as the ground-truth because its degradation is much lighter. Hence, this restoration branch learns to reduce noise and restore the lost information from the reflectance map of the corresponding normal-light image.

Fig. 3 shows the detailed structure of our joint learning network. Similar to LISU-decomp, LISU-joint is also a U-shaped structure, and it has more convolutional layers to extract features. Specifically, LISU-joint takes the three-channel reflectance map and one-channel illumination map generated by LISU-decomp as input, then a shared five-layer encoder extracts representative features. The feature maps are decoded by two decoders and output segmentation map and restored reflectance map, respectively. The two decoders are not isolated because we concatenate the features of decoder layers together. The purpose is to enable the network to make use of the features across different tasks so as to promote a close joint learning of reflectance restoration and semantic segmentation. This is a typical multi-task learning structure used in many papers (Baslamisli et al., 2018; Jiao et al., 2019; Cheng et al., 2017b; Zeng et al., 2019). As the training progresses, the semantic segmentation task benefits from the gradually restored illumination-invariant features. At the same time the segmentation branch also provides semantic information to the restoration branch, and prompts it to produce better restoration at boundaries.

Although we do not have the ground-truth of reflectance maps for

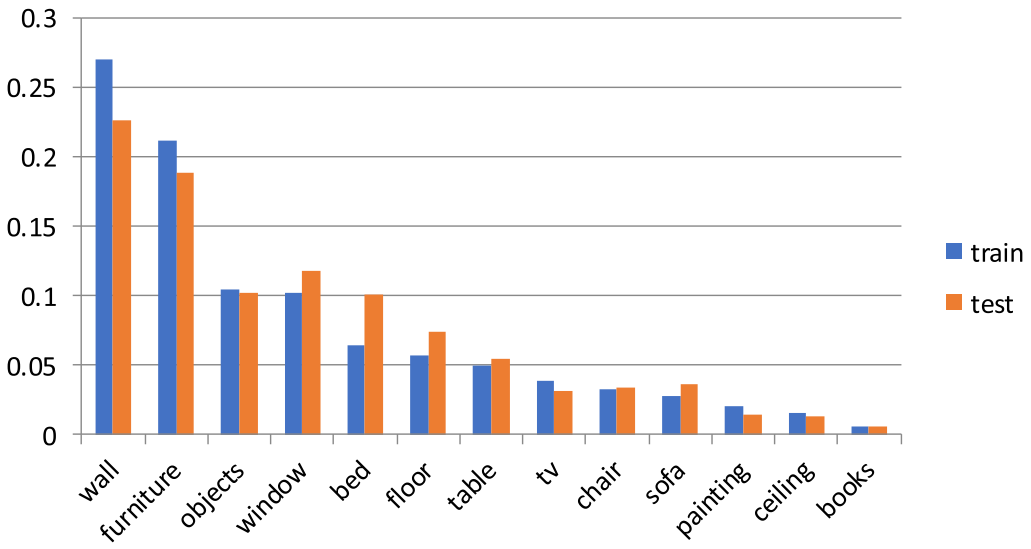


**Fig. 5.** The categories distribution of LLRGBD-real training and test sets. Horizon axis shows the semantic labels, and vertical axis shows the corresponding proportion of pixels.

the restoration task, we can use the reflectance map of the normal-light image as a guide because it is less affected by degradation and it has better quality and more details. We follow (Zhang et al., 2019) and use the following loss function to guide the restoration of reflectance maps:

$$L_{restore} = \| \overline{R}(I_l) - R(I_n) \|_2^2 - SSIM\left(R(I_l), R(I_n)\right) + \| \nabla\overline{R}(I_l) - \nabla R(I_n) \|_2^2, \tag{7}$$

where $\overline{R}(I_l)$ denotes the restored reflectance map, and $\| \cdot \|_2^2$ is mean square error (MSE reconstruction loss). $SSIM(\cdot,\cdot)$ compares the structural similarity (Wang et al., 2004) of two reflectance components. Optimizing the last term reduces the texture differences between the restored reflectance map and the reflectance map of the normal-light image. The cross-entropy loss is used to train the segmentation branch:

$$L_{ce} = -\frac{1}{n}\sum_i \sum_{c \in M} log\left(p_i^c\right), \tag{8}$$

where $p_i^c$ represents the probability of predicting a pixel $i$ belonging to category $c$, and $M$ is the defined class set. The combined loss function for LISU-joint is:

$$L_{joint} = L_{restore} + L_{ce}. \tag{9}$$

## 4. A new data set to study low-light indoor scenes

To study the specific segmentation task in low-light indoor scenes, we present the newly collected data set, which is called LLRGBD. It consists of one large-scale synthetic data set LLRGBD-synthetic and one small-scale real data set called LLRGBD-real. We also provide the corresponding depth maps for investigating other potential computer vision tasks. Fig. 4 shows some sample images in LLRGBD.

### 4.1. LLRGBD-synthetic

LLRGBD-synthetic contains photo-realistic low-/normal-light image pairs, and these images are rendered using the Opposite Renderer (Pedersen, 2013), which was developed based upon the Nvidia OptiX ray-tracing engine. Any rendering engine can be used for scene rendering, such as Blender or UNREAL engine. We choose Opposite Renderer in this paper because it is easy to use, and McCormac et al. (2017) also used this engine and provided some useful learning documents. We modify the source code of the rendering tool so that we can put a white luminous sphere on the camera, and it is the only light source to illuminate scenes. In this way we are able to simulate low-light indoor environments.

When rendering starts, the rendering engine first randomly selects one of 57 pre-defined layouts including bedroom, kitchen, bathroom, living room, and office. Random kinds of objects related to the scene are loaded from the ShapeNet library (Chang et al., 2015), scaled to random sizes and placed at random positions in the scene. We build a high-definition texture library, and random texture maps are assigned to the objects according to their categories. The engine is developed based on real physical rules, hence it can avoid objects from appearing in inappropriate positions. Then the camera moves in the generated scene, and we save the random motion trajectory. Once all parameters related to the scene are determined we set an orb on the camera as a point light source, and control its intensity and radius to render low-light and normal-light RGB images. Besides, we obtain the corresponding depth map via the z-buffer of OpenGL. We annotate the RGB images with 13 classes (Couprie et al., 2013).

We use a Nvidia Titan XP graphical card and each rendering takes two to three minutes. A total of $29K \times 2$ images with a resolution $640 \times 480$ are rendered. We randomly divide LLRGBD-synthetic into training set and test set according to the proportion of 90%-10%.

### 4.2. LLRGBD-real

We collect real images with an Intel RealSense D435i depth camera, a portable RGB-D camera. The traditional method collected low-light images by changing the shutter speed and ISO (Chen et al., 2018; Guo et al., 2016; Wei et al., 2018). However, this method can not simulate the dark scene illuminated by a point light source. When we take low-light images, we turn off all indoor lights to ensure that the room is completely dark and only use one LED light for illumination. Our purpose is to simulate the lighting that UAVs use to explore real dark environments. The LED light we use has a color temperature of $5500 \pm 200$ K, and illuminance of about 800 lm. We take normal-light images by turning on all white lights in the room.

We take images in 32 indoor scenes, including bedroom, kitchen, bathroom, living room, and office. Finally, 515 pairs of low-/normal-light images at $640 \times 480$ resolution are captured, and we manually annotate them using the same 13 classes mentioned above. When we take RGB images, the corresponding depth maps are also collected and aligned to the RGB images using the script provided by the RealSense

**Table 1**
A comparison of representative 3D indoor scene understanding data sets and low-light image data sets. The images provided in Ren et al. (2019) are retouched by trained photographers using image editing software.

| | NYU-Dv2 (Silberman et al., 2012) | SUN RGBD (Song et al., 2015) | SceneNet (Handa et al., 2016) | SceneNet RGBD (McCormac et al., 2017) | LOL (Wei et al., 2018) | LIME (Guo et al., 2016) | Ren et al. (2019) | LLRGBD-real (ours) | LLRGBD-synthetic (ours) |
|---|---|---|---|---|---|---|---|---|---|
| Scene understanding | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ |
| Low-light images | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Illumination settings | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 |
| Number of layouts | 464 | - | 57 | 57 | - | - | - | 32 | 57 |
| Number of images | 1449 | 10 K | 10 K | 5 M | 500 | 10 | 336 | 515 | 29 K |
| Method of design | Real | Real | Rendered | Rendered | Real | Real | Retouched | Real | Rendered |

D435i Development Kit. The depth maps are then post-processed for better smoothness using the algorithm proposed by Levin et al. (2004).

To split the data set we put all the images belonging to the same layout class (kitchen, bedroom, living room, etc.) in the same folder. Then, we randomly select images for training and testing from each folder. Our data sampling strategy ensures that the class distribution of each test set is similar to the class distribution of the corresponding training set. We use 415 image pairs as the training set and 100 image pairs as the test set. Fig. 5 shows the categories distribution of training and test sets on LLRGBD-real.

We compare LLRGBD with some representative RGB-D indoor data sets and low-light image data sets in Table 1. There is a recent data set "See-In-Dark" (Chen et al., 2018) that also provides pairs of low/normal-light data. We do not include it in this table because that data set is focused on enhancing raw sensor data. We can find that the available low-light image data sets are still insufficient, hence our data set plays a role to study semantic segmentation in low-light indoor scenes.

## 5. Evaluation metrics and implementation

### 5.1. Metrics

To compare the proposed LISU with other methods we use three metrics to evaluate the segmentation.

- Overall Accuracy (OA) computes the proportion of pixels that have been correctly classified, and it is defined as:

$$OA = \frac{\sum_{i=0}^{C-1} p_{ii}}{\sum_{i=0}^{C-1}\sum_{j=0}^{C-1} p_{ij}}, \tag{10}$$

where $C$ is the number of predefined categories, including background class. $P_{ii}$ denotes the pixels being predicted correctly. $P_{ij}$ are the pixels belonging to class $i$ but being classified as class $j$.

- Mean accuracy (mAcc.) computes the average of all pixel accuracy over all the categories. It is defined as:

$$mAcc. = \frac{1}{C}\sum_{i=0}^{C-1} \frac{p_{ii}}{\sum_{j=0}^{C-1} p_{ij}}. \tag{11}$$

- Mean intersection over Union (mIoU) calculates the average of all the IoUs over all the categories. IoU is the intersecting part between the predicted pixels and the true labels divided by the union between the predicted pixels and the true labels. It is defined as:

$$mIoU = \frac{1}{C}\sum_{i=0}^{C-1} \frac{p_{ii}}{\sum_{j=0}^{C-1} p_{ij} + \sum_{j=0}^{C-1} p_{ji} - p_{ii}}. \tag{12}$$

This paper studies the semantic segmentation task, but we also want to evaluate if the joint learning improves the task of reflectance restoration. We use the structural similarity index measure (SSIM) (Wang et al., 2004) as metric. We first use Gamma transformation to adjust the illumination map of a low-light image:

$$\overline{S}(I_l) = S(I_l)^{\gamma}, \tag{13}$$

where $\gamma$ is pre-defined and is set to 0.1. Then we obtain the enhanced image:

$$\overline{I}_l = \overline{R}(I_l)\cdot\overline{S}(I_l). \tag{14}$$

**Table 2**
Comparison of accuracy of direct segmentation of low-light images. The results of the first three lines are both trained and tested on the synthetic images, while the results of the last three lines are trained and tested on the real images.

| Data set | Method | OA | mAcc. | mIoU |
|---|---|---|---|---|
| LLRGBD-synthetic | SegNet | 73.4 | 33.0 | 25.6 |
| | U-Net | 77.8 | 43.1 | 33.2 |
| | LISU-seg | **82.3** | **49.0** | **39.5** |
| LLRGBD-real | SegNet | 42.8 | 34.2 | 22.4 |
| | U-Net | 54.6 | 47.4 | 32.8 |
| | LISU-seg | **59.0** | **50.0** | **36.1** |

**Table 3**
Comparison of segmentation accuracy of degraded reflectance maps.

| Data set | Method | OA | mAcc. | mIoU |
|---|---|---|---|---|
| LLRGBD-synthetic | SegNet | 80.3 | 45.1 | 35.9 |
| | U-Net | 78.7 | 45.5 | 35.6 |
| | LISU-seg | **82.5** | **49.9** | **40.3** |
| LLRGBD-real | SegNet | 48.2 | 36.8 | 26.2 |
| | U-Net | 59.3 | 52.1 | 38.1 |
| | LISU-seg | **60.1** | **52.6** | **38.9** |

We compute $SSIM(\overline{I}_l, I_n)$, and a higher value means that the restored reflectance is closer to the reflectance map of the normal-light image.

### 5.2. Implementation details

We develop the proposed LISU framework using PyTorch on a Linux system with Nvidia Titan XP GPU as the graphics card. In the training phase, we resize all the images to $320 \times 240$ and we do not apply any data augmentation. The batch size is set to 12, and the optimizer is Adam solver with $(\beta_1, \beta_2) = (0.95, 0.999)$. We set the initial learning rate to 0.001 with polynomial decay with power $p = 0.9$. We train 50 epochs on LLRGBD-synthetic and 200 epochs on LLRGBD-real. The weights in Eq. 6 are set as $\lambda_1 = 0.01, \lambda_2 = 0.1$ and $\lambda_3 = 0.5$. We study the influence of different combinations of weights on segmentation results and we present this part in Section 6.4.

## 6. Experiments and discussion

### 6.1. Segmentation with the baseline model: LISU-seg

We use the segmentation branch of LISU-joint as the baseline model **LISU-seg**. Particularly, it only has one decoder to output the segmentation map, and it does not have the reflectance restoration features. The baseline model takes low-light images as input. We compare our baseline model with SegNet (Badrinarayanan et al., 2017) and U-Net (Ronneberger et al., 2015), which are both encoder-decoder structures designed for semantic segmentation task. The results are shown in Table 2 and we can find our baseline model outperforms SegNet and U-Net on both synthetic and real data sets. SegNet uses max pooling layers to reduce the size of the feature map, but at the same time some spatial information is lost. U-Net adopts max pooling as well, but it also introduces the long-skip connection which helps to preserve the information from encoder layers. In the baseline model LISU-seg, we also use long-skip connection, but we do not use max pooling. Instead, we control the strides of convolutional layers to reduce the sizes of feature maps.

### 6.2. Segmentation using degraded reflectance maps

The illumination-invariant reflectance maps of low-light images should be helpful to the segmentation task. In this experiment we explore if training on the degraded reflectance maps generated by LISU-decomp can improve the segmentation accuracy. We first use low-light images to train the unsupervised decomposition network LISU-decomp. Then, for each batch of low-light images the trained LISU-decomp output three-channel reflectance maps and one-channel illumination maps. We input these reflectance maps to train the segmentation networks. As shown in Table 3, although the reflectance maps generated at this stage degrade heavily, they make all three networks achieve better accuracy than segmenting on the original low-light images.

**Table 4**
Evaluation of LISU. We copy the results of the baseline model LISU-seg from Table 2 for easier comparison.

| Data set | Method | OA | mAcc. | mIoU |
|---|---|---|---|---|
| LLRGBD-synthetic | LISU-seg | 82.3 | 49.0 | 39.5 |
| | LISU | **84.5** | **52.3** | **43.4** |
| LLRGBD-real | LISU-seg | 59.0 | 50.0 | 36.1 |
| | LISU | **67.3** | **61.2** | **47.6** |

**Table 5**
Results of ten repeated experiments. Both training and testing were performed on our real data.

| No. | OA | mAcc. | mIoU |
|---|---|---|---|
| 1 | 67.3 | 61.2 | 47.6 |
| 2 | 67.8 | 61.6 | 47.6 |
| 3 | 64.8 | 60.4 | 47.2 |
| 4 | 67.4 | 59.7 | 46.8 |
| 5 | 66.0 | 65.9 | 46.0 |
| 6 | 69.0 | 61.4 | 48.4 |
| 7 | 65.9 | 58.4 | 45.7 |
| 8 | 67.7 | 62.1 | 49.3 |
| 9 | 66.1 | 60.1 | 46.6 |
| 10 | 68.7 | 59.3 | 45.9 |
| mean | 67.07 | 61.01 | 47.11 |
| standard deviation | 1.34 | 2.06 | 1.15 |

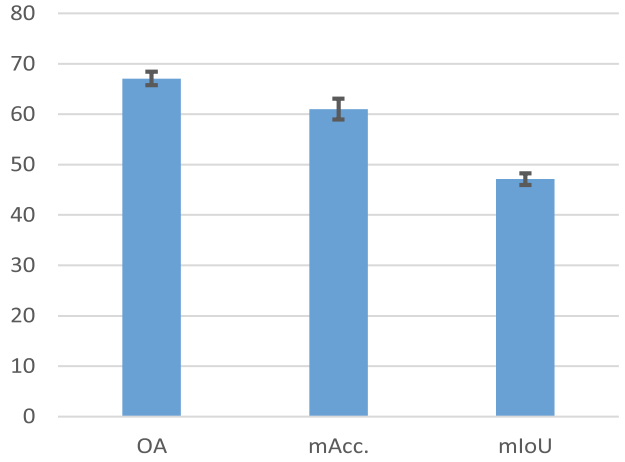


**Fig. 6.** Average values of metrics of 10 experiments with error bars.

### 6.3. Segmentation with the proposed LISU framework

In this experiment we evaluate the effectiveness of the proposed LISU framework, which contains two sub-networks, LISU-decomp and LISU-joint. Both sub-networks update their parameters independently. As shown in Table 4, compared with our baseline model LISU substantially improves the mIoU from 36.1 to 47.6 on the LLRGBD-real data set. When training on the small low-light data set, the semantic segmentation task benefits from the restored reflectance features. The improvement also reflects in our synthetic data set that the mIoU increases from 39.5 to 43.4. The reason why the improvement is not as great as in the real data

**Table 6**
Study of the influence of weights of Eq. 6 and joint learning on segmentation and reflectance restoration. The best and the second best results are highlighted in **bold** and underlined, respectively.

| No. | Weights of Eq. 6 | | | LISU (with LISU-semi-joint) | | | | LISU (with LISU-joint) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | OA | mAcc. | mIoU | SSIM | OA | mAcc. | mIoU | SSIM |
| 1 | 0.01 | 0.1 | 0.5 | **67.5** | 59.1 | **46.1** | **0.6846** | 67.3 | **61.2** | **47.6** | **0.6860** |
| 2 | 0.1 | 0.1 | 0.5 | 65.6 | 58.6 | 44.9 | 0.6807 | 65.7 | 58.9 | 45.8 | 0.6816 |
| 3 | 0.5 | 0.1 | 0.5 | 62.2 | 54.7 | 41.0 | 0.6717 | 64.1 | 55.8 | 42.3 | 0.6735 |
| 4 | 0.01 | 0.5 | 0.5 | 64.9 | 56.5 | 43.0 | 0.6716 | 65.0 | 57.9 | 44.6 | 0.6746 |
| 5 | 0.01 | 0.01 | 0.5 | 66.6 | **59.8** | 45.1 | 0.6654 | **68.7** | 60.0 | 47.5 | 0.6668 |
| 6 | 0.01 | 0.01 | 0.1 | 66.6 | 57.9 | 44.5 | 0.6634 | 67.1 | 59.7 | 47.1 | 0.6658 |
| 7 | 0.01 | 0.01 | 0.01 | 64.1 | 56.8 | 42.9 | 0.6557 | 65.2 | 55.9 | 43.1 | 0.6553 |

set is that the synthetic data are rendered by a camera moving according to a trajectory at a fixed frame rate, *i.e.,* when the camera view changes we can get the photos of the same object under different illumination. These large numbers of sequential images enable us to train a network that is robust to illumination changes.

To further verify our approach's generalization capability we repeat the training/testing a total of 10 times. For each training/testing, the whole data set is randomly split into a training set and a testing set as described in Section 4.2. The results of our ten experiments are shown in Table 5, and we show the average values of metrics of 10 experiments with error bars in Fig. 6. The results show that our framework has kept good generalization ability in 10 experiments. When there is not enough data, it is feasible to use small data sets to train a deep model. As shown by Zuo et al., a small data set can still allow to reliably train and test networks (Zuo and Drummond, 2017). Our experiments are repeated by reshuffling the samples and we always observed normal training curves, without any evidence of over-fitting. We finally chose experiment No.1 and publish its corresponding training set and test set because each result of this experiment is closest to the average value of each evaluation metric.

### 6.4. Ablation study of the joint learning branches

The decomposition network plays a vital role in our method, as we use its output for subsequent learning. In this experiment we evaluate the influence of weights in Eq. 6 on joint learning. At the same time we also explore how the linked features from two tasks affect the final results.

Our study of weights is based on a grid search. However, we cannot traverse all the coefficient scales because there are infinite combinations. We observed training losses and found that when we set $\lambda_1$,$\lambda_2$, or $\lambda_3$ to be greater than or close to 1, the decomposition network did not converge. Therefore, we selected three values of 0.01, 0.1, 0.5 for each weight and studied the influence of different combinations of weights. However, we do not need to run experiments 27 times. By fixing two of the weights and observing the influence of the third weight, we use seven weight combinations. For each combination we train two joint learning networks with different feature connections strategies. One is the default LISU-joint that the features from decoders of two tasks are concatenated together. The other is that the features from the segmentation branch are not linked to the reflectance restoration branch, that is, the LISU-joint structure in Fig. 3 but without copying the yellow features from the segmentation branch. We call this variant LISU-semi-joint.

As shown in the first row to the third row of Table 6, larger $\lambda_1$ yields worse segmentation. The larger $\lambda_1$ forces the decomposition network to generate closer reflectance maps from paired images. Yet the degraded reflectance map of a low-light image has information loss, and simply increasing $\lambda_1$ will make the decomposition network output a reflectance map without detailed textures. The $\lambda_2$ controls the illumination map to be close to the maximum values of RGB channels of the original low-light image. Higher segmentation accuracy can be achieved when $\lambda_2$ is 0.1 or 0.01. We finally choose the first combination of weights as the final parameters ($\lambda_1 = 0.01, \lambda_2 = 0.1, \lambda_3 = 0.5$), because this setting can achieve decent segmentation accuracy and the highest SSIM score. When we use the same weights setting, LISU with LISU-joint has higher segmentation accuracy and SSIM score than LISU with LISU-semi-joint. This demonstrates that semantic information helps to restore reflectance, and when we have better reflectance features, the segmentation accuracy is further improved.

**Table 7**
Evaluation of the effectiveness of pre-training on LLRGBD-synthetic.

| Data set | Method | OA | mAcc. | mIoU |
|---|---|---|---|---|
| LLRGBD-real | LISU (pre-trained) | 72.3 | 68.2 | 54.8 |

### 6.5. Pre-training on LLRGBD-synthetic

In order to make up for the shortage of real data sets, some semantic segmentation models are pre-trained on synthetic data (Handa et al., 2016; McCormac et al., 2017). In this experiment we also study if our rendered data set is helpful to improve the segmentation accuracy as pre-training data. We use LLRGBD-synthetic to pre-train LISU for 50 epochs. Then, we freeze the encoders of LISU-decomp and LISU-joint, and fine-tune the model by training on LISU-real and updating the decoders' parameters. Table 7 shows that the pre-training improves the mIoU from 47.6 to 54.8, which means our synthetic low-light data set is photo-realistic, and it provides more training images to enhance the learning ability of the network when the real low-light data set is insufficient.

### 6.6. Segmentation with DeepLab v3 + and its variants

In this experiment we evaluate the performance of the state-of-the-art semantic segmentation network DeepLab v3+ (Chen et al., 2018) (DLv3p) on our real data set. It is shown in the yellow box (without green feature maps) in Fig. 7, and it is also an encoder-decoder structure and uses ResNet50 as the backbone (He et al., 2016). It uses the spatial pyramid pooling (SPP) module to capture multi-scale information. The feature maps from the encoder are up-sampled by 4 times and concatenated with the feature maps from the second layers of the backbone.

In addition, we have added an extra decoder to its original structure, so that it can restore reflectance maps. Like our LISU-joint, the yellow feature maps from segmentation branch and the green feature maps from reflectance restoration branch are concatenated together for joint learning, a variant we call **DLv3p-joint**. Similar to the experiment in Section 6.4 we also use another variant called **DLv3p-semi-joint**, in which the feature maps of the restoration branch are not copied to the segmentation branch. We train DLv3p and its variants on LLRGBD-real. Note that for the modified structures DLv3p-joint and DLv3p-semi-joint, we first train LISU-decomp for low-light image decomposition, and then we use the output of LISU-decomp to train the variants with joint learning. We come across an overfitting problem when training DLv3p, so we adopt early-stopping strategy. The reason is that the segmentation task on this data set is too simple for such a complex model with a large

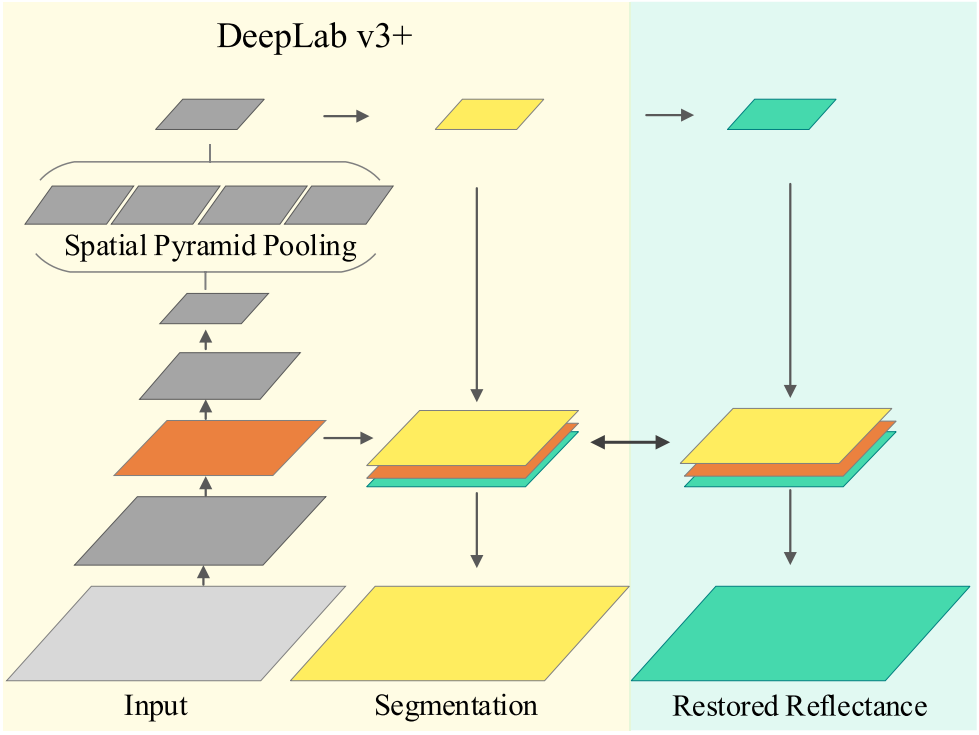


**Fig. 7.** A modified DeepLab v3 + that jointly learns to restore reflectance.

**Table 8**
Comparison of segmentation accuracy using DeepLab v3 + and its variants.

| Method | OA | mAcc. | mIoU |
|---|---|---|---|
| DLv3p (Chen et al., 2018) | 54.9 | 46.9 | 33.4 |
| Dlv3p-semi-joint | 67.2 | 60.4 | 47.2 |
| DLv3p-joint | **68.4** | **62.0** | **49.2** |

amount of parameters. It can perfectly fit the data in the training set, but it loses its generalization ability in the test set. The results shown in Table 8 demonstrate that our proposed approach can be easily integrated with the encoder-decoder structure and achieves better results. Compared with DLv3p-semi-joint, the strategy of combining features from both encoders (DLv3p-joint) improves the accuracy by 2.0% (for mIoU 49.2 versus 47.2). Note that only one decoder layer is involved in the feature concatenation, which once again proves the validity of our method. We list the detailed quantitative comparison of some networks in Table 9. We do not include any comparisons of traditional methods (Alshammari et al., 2018; Maddern et al., 2014; Upcroft et al., 2014) because only (Maddern et al., 2014) provides the source code. However, this method needs to know the peak spectral response of the RGB sensor, which is not provided by Intel RealSense D435i. Fig. 8 shows the qualitative results. Although normal-light images are not needed when inferring the model, they are still shown here for better visual comparison.

### 6.7. Discussion on failure cases

When taking low-light images, some materials such as glass and ceramic tiles will reflect LED light and produce white spots, as shown in the first row of Fig. 9. These spots can be seen as overexposure and information is lost. Therefore, they bring challenges to the decomposition network and affect the final segmentation results. Although these spots account for only a small part of the data set, it is still useful if future work attempts to design a better image decomposition network to eliminate the influence of these white spots. Besides, increasing the accuracy on boundaries will be the focus of future work. As shown in the second row of Fig. 9, the boundary between the flowerpot and the sofa is wrongly segmented due to the similar colors and textures. Since we also provide depth map in our data sets, future work can utilize geometric information to improve the segmentation on boundaries.

## 7. Conclusion

In this paper, we have studied semantic segmentation in low-light indoor environments. We find that existing CNN-based networks, even the state-of-the-art DeepLab v3+ (Chen et al., 2018), cannot deal with illumination changes effectively. Hence, we present a novel end-to-end trainable CNN framework that takes advantage of the illumination-invariant features for low-light indoor scene segmentation. We show that the joint learning of reflectance restoration and semantic segmentation benefits both tasks, and the segmentation task benefits most from

**Table 9**
Quantitative comparison on each class of the LLRGBD-real data set. The IoUs are shown for evaluation, and the best results are highlighted in **bold**.

| Method | bed | books | ceiling | chair | floor | furniture | objects | painting | sofa | table | tv | wall | window | mAcc. | mIoU | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LISU-seg (baseline) | 46.5 | 12.3 | 7.8 | 24.1 | 45.5 | 34.4 | 33.6 | 41.1 | 35.2 | 32.1 | 62.6 | 60.4 | 33.7 | 59.0 | 50.0 | 36.1 |
| DLv3p (Chen et al., 2018) | 46.4 | 29.1 | 12.5 | 20.5 | 42.4 | 30.0 | 27.0 | 37.5 | 18.9 | 26.0 | 56.8 | 55.6 | 31.3 | 54.9 | 46.9 | 33.4 |
| LISU | 57.8 | 45.1 | 29.6 | 34.3 | 52.0 | 41.6 | 39.8 | 47.3 | 44.3 | 36.5 | 70.6 | 70.7 | 48.6 | 67.3 | 61.2 | 47.6 |
| DLv3p-joint | **64.8** | 55.6 | 22.6 | 31.0 | 54.0 | 39.2 | 41.5 | 56.9 | 42.0 | **40.3** | 71.4 | 70.1 | 50.8 | 68.4 | 62.0 | 49.2 |
| LISU (pre-trained) | 62.2 | **63.0** | **42.8** | **36.3** | **59.0** | **48.8** | **47.2** | **58.8** | **56.1** | 36.0 | **71.8** | **74.9** | **55.2** | **72.3** | **68.2** | **54.8** |

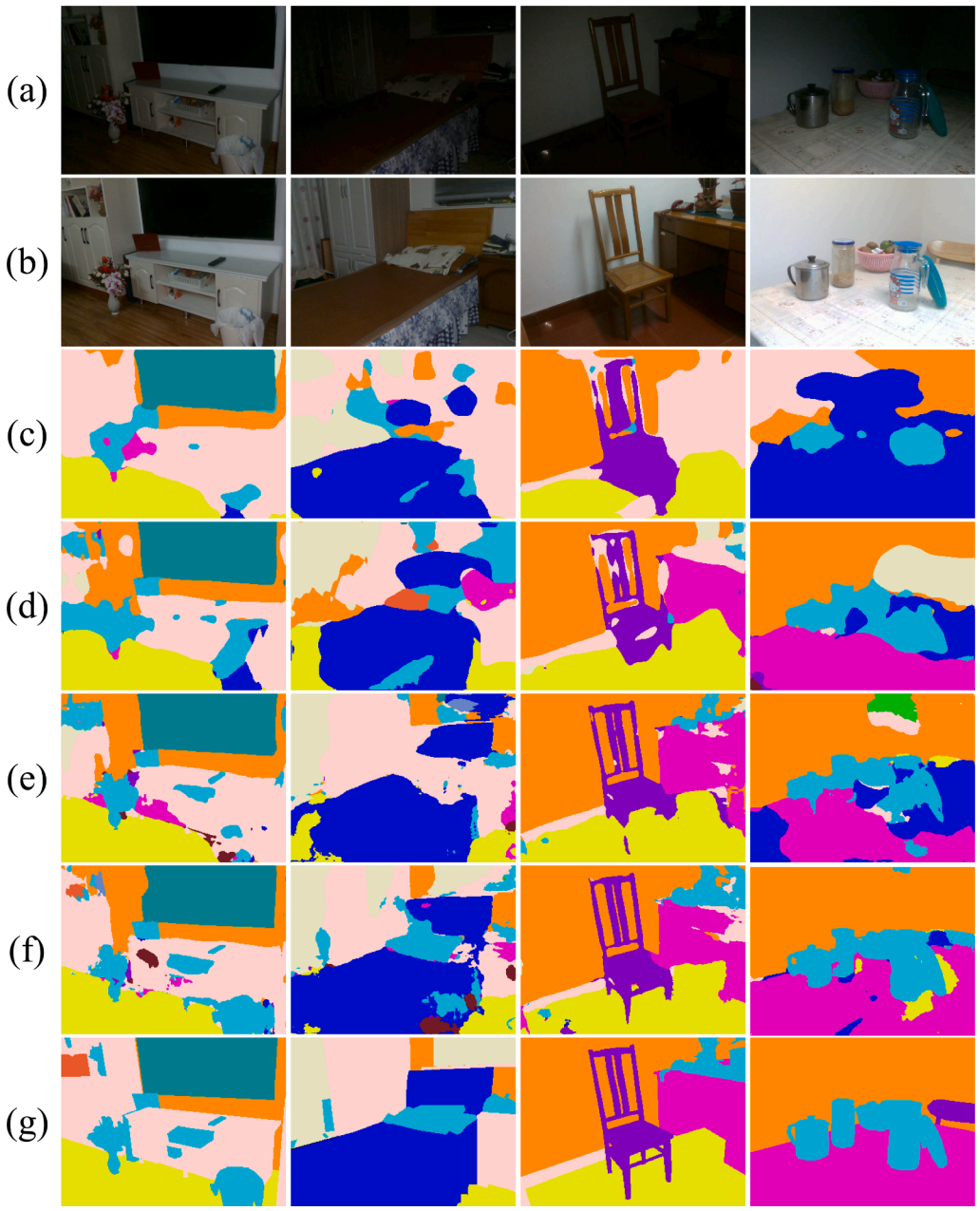


**Fig. 8.** Qualitative results on LLRGBD-real. (a) is the input low-light images, and (b) shows the corresponding normal-light images; (c): DLv3p (Chen et al., 2018); (d): LISU-seg; (e) LISU; (f) LISU (pre-trained on LLRGBD-synthetic); (g) Ground-truth.

(a) low-light image (b) degraded R (c) restored R (d) output (e) semantic label

**Fig. 9.** Failure cases. The first row shows the white spot caused by reflective materials. The second row shows the failure segmentation on boundaries. (a) is the input low-light image; (b) is the reflectance map output by LISU-decomp; (c) and (d) are the restored reflectance and segmentation map output by LISU-joint, respectively; (e) is the segmentation ground-truth. Red rectangles denote the regions of interest.

the joint learning network. The illumination-invariant features of reflectance restoration branch are helpful for segmentation. Also, the semantic information helps to generate sharper reflectance maps, and then better reflectance features further improve the segmentation results. Besides, the proposed photo-realistic synthetic data set and real data set are complementary to the research of indoor scene understanding, especially in low-light environments. Experimental results demonstrate that our approach achieves favorable segmentation performance. The potential applications of this research include UAVs or ground robots understanding low-light scenes and navigating autonomously in them. As part of the INGENIOUS project we will extend our algorithm and segment low-light scenes partially damaged by earthquakes, so as to improve first responders' situational awareness at the disaster sites.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme and the Korean Government under Grant Agreement No 833435. Content reflects only the authors' view and European Commission is not responsible for any use that may be made of the information it contains.

## References

Adachi, M., Shatari, S., Miyamoto, R., 2019. Visual navigation using a webcam based on semantic segmentation for indoor robots. In: IEEE SITIS, 2019, pp. 15–21.
Alshammari, N., Akcay, S., Breckon, T.P., 2018. On the impact of illumination-invariant image pre-transformation for contemporary automotive semantic scene understanding. in: Intelligent Vehicles Symposium, 2018.
Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE TPAMI 39 (12), 2481–2495.
Baslamisli, A.S., Groenestege, T.T., Das, P., Le, H.-A., Karaoglu, S., Gevers, T., 2018. Joint learning of intrinsic images and semantic segmentation. In: ECCV.
Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F., 2015. Shapenet: An information-rich 3d model repository, arXiv preprint arXiv:1512.03012.
Chen, C., Chen, Q., Xu, J., Koltun, V., 2018. Learning to see in the dark. In: CVPR.
Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV, 2018.
Cheng, Y., Cai, R., Li, Z., Zhao, X., Huang, K., 2017a. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In: CVPR, 2017.
Cheng, J., Tsai, Y.-H., Wang, S., Yang, M.-H., 2017b. Segflow: Joint learning for video object segmentation and optical flow. In: ICCV, 2017, pp. 686–695.
Cho, S.W., Baek, N.R., Koo, J.H., Arsalan, M., Park, K.R., 2020. Semantic segmentation with low light images by modified cyclegan-based image enhancement. IEEE Access 8, 93561–93585.
Couprie, C., Farabet, C., Najman, L., LeCun, Y., 2013. Indoor semantic segmentation using depth information. In: ICLR, 2013.
Dai, D., Van Gool, L., 2018. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In: ITSC.
Dai, J., He, K., Sun, J., 2016. Instance-aware semantic segmentation via multi-task network cascades. In: CVPR, 2016.
Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D., 018. Revisiting deep intrinsic image decompositions. In: CVPR.
Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia-Rodriguez, J., 2017. A review on deep learning techniques applied to semantic segmentation, arXiv preprint arXiv:1704.06857.
Giernacki, W., Skwierczyński, M., Witwicki, W., Wroński, P., Kozierski, P., 2017. Crazyflie 2.0 quadrotor as a platform for research and education in robotics and control engineering. In: MMAR, IEEE, 2017, pp. 37–42.
Guo, X., Li, Y., Ling, H., 2016. Lime: Low-light image enhancement via illumination map estimation. IEEE TIP 26 (2), 982–993.
Gupta, S., Sangeeta, R., Mishra, R.S., Singal, G., Badal, T., Garg, D., 2020. Corridor segmentation for automatic robot navigation in indoor environment using edge devices. Comput. Netw. 178, 107374.
Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S., Cipolla, R., 2016. Understanding real world indoor scenes with synthetic data. In: CVPR.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: CVPR.
Janner, M., Wu, J., Kulkarni, T.D., Yildirim, I., Tenenbaum, J., 2017. Self-supervised intrinsic image decomposition. In: NeurIPS.
Jiao, J., Wei, Y., Jie, Z., Shi, H., Lau, R.W., Huang, T.S., 2019. Geometry-aware distillation for indoor semantic segmentation. In: CVPR, 2019.
Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS J. Photogramm. Remote Sens 145, 60–77.
Kwon, Y.-S., Lim, H., Jung, E.-J., Yi, B.-J., 2008. Design and motion planning of a two-moduled indoor pipeline inspection robot. In: ICRA, IEEE, 2008, pp. 3998–4004.
Land, E.H., McCann, J.J., 1971. Lightness and retinex theory. J. Opt. Soc. Am. 61 (1), 1–11.
Lau, H.Y., Ko, A., 2007. An immuno robotic system for humanitarian search and rescue (application stream). In: ICARIS. Springer, pp. 191–203.
Levin, A., Lischinski, D., Weiss, Y., 2004. Colorization using optimization. In: SIGGRAPH.
Li, Z., Snavely, N., Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In: ECCV.
Li, F., Zlatanova, S., Koopman, M., Bai, X., Diakité, A., 2018. Universal path planning for an indoor drone. Automation in Construction 95, 275–283.
Lin, G., Milan, A., Shen, C., Reid, I., 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: CVPR.
Liu, Y., Li, Y., You, S., Lu, F., 2020. Unsupervised learning for intrinsic image decomposition from a single image. In: CVPR.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: CVPR.
Lu, Y., Xue, Z., Xia, G.-S., Zhang, L., 2018. A survey on vision-based uav navigation. Geo-spatial information science 21 (1), 21–32.
W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, P. Newman, Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles. In: ICRA, 2014.
J. McCormac, A. Handa, S. Leutenegger, A.J. Davison, Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation?. In: ICCV, 2017.
T. Narihira, M. Maire, S.X. Yu, Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In: ICCV, 2015.
G. Neuhold, T. Ollmann, S. Rota Bulo, P. Kontschieder, The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV, 2017, pp. 4990–4999.
Özaslan, T., Shen, S., Mulgaonkar, Y., Michael, N., Kumar, V., 2015. Inspection of penstocks and featureless tunnel-like environments using micro uavs. In: Field and Service Robotics. Springer, pp. 123–136.
S.-J. Park, K.-S. Hong, S. Lee, Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In: ICCV, 2017.
S.A. Pedersen, Progressive photon mapping on gpus, Master's thesis, Institutt for datateknikk og informasjonsvitenskap (2013).
K. Rematas, T. Ritschel, M. Fritz, E. Gavves, T. Tuytelaars, Deep reflectance maps. In: CVPR, 2016.
Ren, W., Liu, S., Ma, L., Xu, Q., Xu, X., Cao, X., Du, J., Yang, M.-H., 2019. Low-light image enhancement via a deep hybrid network. IEEE TIP 28 (9), 4364–4375.
O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation. In: MICCAI, 2015.
C. Rother, M. Kiefel, L. Zhang, B. Schölkopf, P.V. Gehler, Recovering intrinsic images with a global sparsity prior on reflectance. In: NeurIPS, 2011.
C. Sakaridis, D. Dai, L.V. Gool, Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: ICCV, 2019.
L. Shen, P. Tan, S. Lin, Intrinsic image decomposition with non-local texture cues. In: CVPR, 2008.
N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images. In: ECCV, 2012.
S. Song, S.P. Lichtenberg, J. Xiao, Sun rgb-d: A rgb-d scene understanding benchmark suite. In: CVPR, 2015.
L. Sun, K. Wang, K. Yang, K. Xiang, See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion. In: Security + Defence, 2019.
Tappen, M.F., Freeman, W.T., Adelson, E.H., 2005. Recovering intrinsic images from a single image. IEEE TPAMI 27 (9), 1459–1472.
B. Upcroft, C. McManus, W. Churchill, W. Maddern, P. Newman, Lighting invariant urban street classification. In: ICRA, 2014.
Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE TIP 13 (4), 600–612.
C. Wei, W. Wang, W. Yang, J. Liu, Deep retinex decomposition for low-light enhancement. In: BMVC, 2018.
B. Wu, A. Wan, X. Yue, K. Keutzer, Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In: ICRA, 2018.
C. Xu, K. Wang, K. Yang, R. Cheng, J. Bai, Semantic scene understanding on mobile device with illumination invariance for the visually impaired. In: Security + Defence, 2019.
Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, Joint learning of saliency detection and weakly supervised semantic segmentation. In: ICCV, 2019, pp. 7223–7233.
Zhang, Y., Yang, Q., 2021. A survey on multi-task learning. IEEE TKDE.
Zhang, Y., Zhang, J., Guo, X., 2019. Kindling the darkness: A practical low-light image enhancer. In: ACM MM, 2019.
Zhu, A., Zhang, L., Shen, Y., Ma, Y., Zhao, S., Zhou, Y., 2020. Zero-shot restoration of underexposed images via robust retinex decomposition. In: ICME, 2020.
Zuo, Y., Drummond, T., 2017. Fast residual forests: Rapid ensemble learning for semantic segmentation. In: CoRL, PMLR, 2017, pp. 27–36.