**Disentangle then Parse:**
**Night-time Semantic Segmentation with Illumination Disentanglement**

Zhixiang Wei\(^1\) \* Lin Chen\(^1,2\) \* Tao Tu\(^1\) Pengyang Ling\(^1\) Huaian Chen\(^1\)† Yi Jin\(^1\)†

\(^1\) University of Science and Technology of China \(^2\) Shanghai AI Laboratory

\{zhixiangwei, chlin, tutao9036, lpyang27, anchen\}@mail.ustc.edu.cn, jinyi08@ustc.edu.cn

1. **NightCity-fine**

NightCity-fine is a refined dataset for night-time semantic segmentation, which aims to improve the quality of annotations in both the training and validation sets. This dataset builds upon NightCity [5], which is the largest dataset for night-time segmentation, and Nightcity+ [2], a validation set based on NightCity. Our annotation process begins by comparing the annotations in the NightCity dataset with the original images, which identified a significant number of missing labels and incorrectly labeled regions. To address these issues, we utilized the graphical image annotation tool, Labelme [6], to accurately label the previously missing regions and remove incorrect labels from mislabeled regions.

As a result, we successfully eliminate 4747 mislabeled regions and rectify 14288 missing labels with the appropriate label, as depicted in Fig. 1. In total, we refined 84% of the images in the dataset, and added 14228 shapes to each category, including both things and stuff, as illustrated in Fig. 2. Among them, traffic light and traffic sign have the most significant numbers of 2981 and 2963, respectively. We compared the pixel distributions of labeled regions in NightCity and NightCity-fine, as shown in Fig. 3. As some categories have significantly higher numbers of pixels than others, we presented the distributions in log scale. Overall, our refined dataset has more balanced pixel distributions of all classes than the original dataset, with more labeled regions for traffic sign, traffic light, motorcycle, wall, and pole, among others. A qualitative comparison of NightCity and NightCity-fine dataset can be found in Fig. 4 and Fig. 5.

2. **Study on guidance noise**

We select two guidance noise distributions. The first noise distribution is based on the Gaussian distribution recommended by [3], which is added to generated illumination to prevent the model to produce an identity result. To be specific, we apply a noise \( N \) that follows the standard normal distribution \( N(0, 1) \). Subsequently, to account for the smoothness and value range of illumination, we normalize the noise \( N \) to the range of 0 to 1 and pass it through an average pooling layer with a kernel size of 16 and stride size of 16. Moreover, inspired by the work of [1, 8], we also use the normalized V channel of the input image in the HSV color space as another guidance noise \( V \). Similarly, this noise is fed to a max pooling layer with a kernel size of 16 and a...
3. Model architecture

Disentanglement model. The network architecture includes a stem layer, multiple downsampling convolution layers, several residual blocks, two Swin blocks [4], a pyramid pooling module [7], several upsampling convolution layers, and two output convolution layers. The reflectance is obtained by adding the output and input images. Tab. 2 displays the number of blocks at different depths.

IAParser. The proposed IAParser consists of several components, including a reflectance segmentation model $M_{ref}$ that can be substituted with an existing semantic segmentation network, an illumination segmentation model $M_{ill}$ that adopts a pyramid pooling module architecture [7], a convolution layer $W_{mask}$ that calculates the attention mask, and a convolution layer $W_{cls}$ that transforms the features produced by $M_{ref}$ and $M_{ill}$ into semantic logits.

4. Algorithm

The training procedure of our DTP is summarized in Algorithm 1, which is composed of semantic-oriented disentanglement (SOD) and illumination-aware parser (IAParser). For detailed equations and loss functions, please refer to the main paper.

5. Qualitative results

Fig. 6, 7, and 8 showcase the reflectance and semantic segmentation results produced by our proposed method. Although the limited dataset scale and model parameters resulted in incomplete disentanglement, which led to the presence of redundant lighting-specific components in the reflectance, our approach effectively enhances the model’s ability to parse images and generate superior semantic segmentation results. The observed improvement in both visual

Table 1. Ablation Studies on the effects of different guide illumination. Random is random generated smooth noise, and Max is the maximum channel of a random picture. The best scores are indicated in **bold**.

<table>
<thead>
<tr>
<th>Guidance Noise</th>
<th>None</th>
<th>N</th>
<th>V</th>
<th>N + V</th>
</tr>
</thead>
<tbody>
<tr>
<td>mIoU(%)</td>
<td>61.6</td>
<td>63.9</td>
<td>62.7</td>
<td>64.2</td>
</tr>
</tbody>
</table>

Table 2. Detailed architecture of disentanglement model with different depths.

<table>
<thead>
<tr>
<th>Disentangle model</th>
<th>Small</th>
<th>Base</th>
<th>Large</th>
<th>Huge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downsampling blocks</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Residual blocks</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Channels of Residual</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>Upsampling blocks</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Parameters</td>
<td>0.2M</td>
<td>1.5M</td>
<td>5.9M</td>
<td>26.8M</td>
</tr>
</tbody>
</table>

Figure 3. Number of labels in each category before and after refining.
**Algorithm 1:** Training process of DTP.

**Input:** disentanglement models: $M_{dis}$; reflectance segmentation model: $M_{ref}$; illumination segmentation mode: $M_{ill}$; convolution layers: $W_{mask}, W_{cls}$; maximum iteration $T$.

**Output:** finale network consists of $M_{dis}, M_{ref}, M_{ill}, W_{mask}, W_{cls}$.

**for** $t ← 1$ **to** $T$ **do**

- Get batch data: $(X_j, Y_j, X_k, Y_k)$
- $R_j, I_j = M_{dis}(X_j)$
- $R_k, I_k = M_{dis}(X_k)$
- Get $I'_j, I'_k$ by Eq. (3)
- Get $R^s_j, I^s_j, R^s_k, I^s_k$ by Eq. (4)
- Calculate $L_{disentangle}$ by Eq. (5)

**for** $R, I, X$ in $(R_j, I_j, X_j), (R_k, I_k, X_k), (R^s_j, I^s_j, R_j ⊙ I'_j), (R^s_k, I^s_k, R_k ⊙ I'_k)$ **do**

- Calculate $L_{retinex}(R, I, X)$ by Eq. (5)
- Calculate $L_{smooth}(R, I)$ by Eq. (6)

**for** $R, I, Y$ in $(R_j, I_j, Y_j), (R_k, I_k, Y_k), (R^s_j, I^s_j, Y_j), (R^s_k, I^s_k, Y_k)$ **do**

- $F_{ill} = M_{ill}(I)$
- $F_{ref} = M_{ref}(R)$
- Calculate $A_{mask}$ by Eq. (9)
- Calculate $Ŷ$ by Eq. (10)
- Calculate $L_{segill}$ by Eq. (11)
- Calculate $L_{seg}$ by Eq. (12)

Optimize network

quality and mIoU metrics (refer to the main paper) supports the effectiveness and competitiveness of our method.
Figure 4. Qualitative Comparison of NightCity dataset with NightCity-fine dataset.
Figure 5. Qualitative Comparison of NightCity dataset with NightCity-fine dataset.
Figure 6. Qualitative Results of the reflectance and semantic segmentation outcomes produced by our proposed method.
Figure 7. Qualitative Results of the reflectance and semantic segmentation outcomes produced by our proposed method.
Figure 8. Qualitative Results of the reflectance and semantic segmentation outcomes produced by our proposed method.
References


