L-CAD: Language-based Colorization with Any-level Descriptions using Diffusion Priors

Zheng Chang[1#] Shuchen Weng[2,3#] Peixuan Zhang[1] Yu Li[4] Si Li[*1] Boxin Shi[2,3]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[4]International Digital Economy Academy

Outline

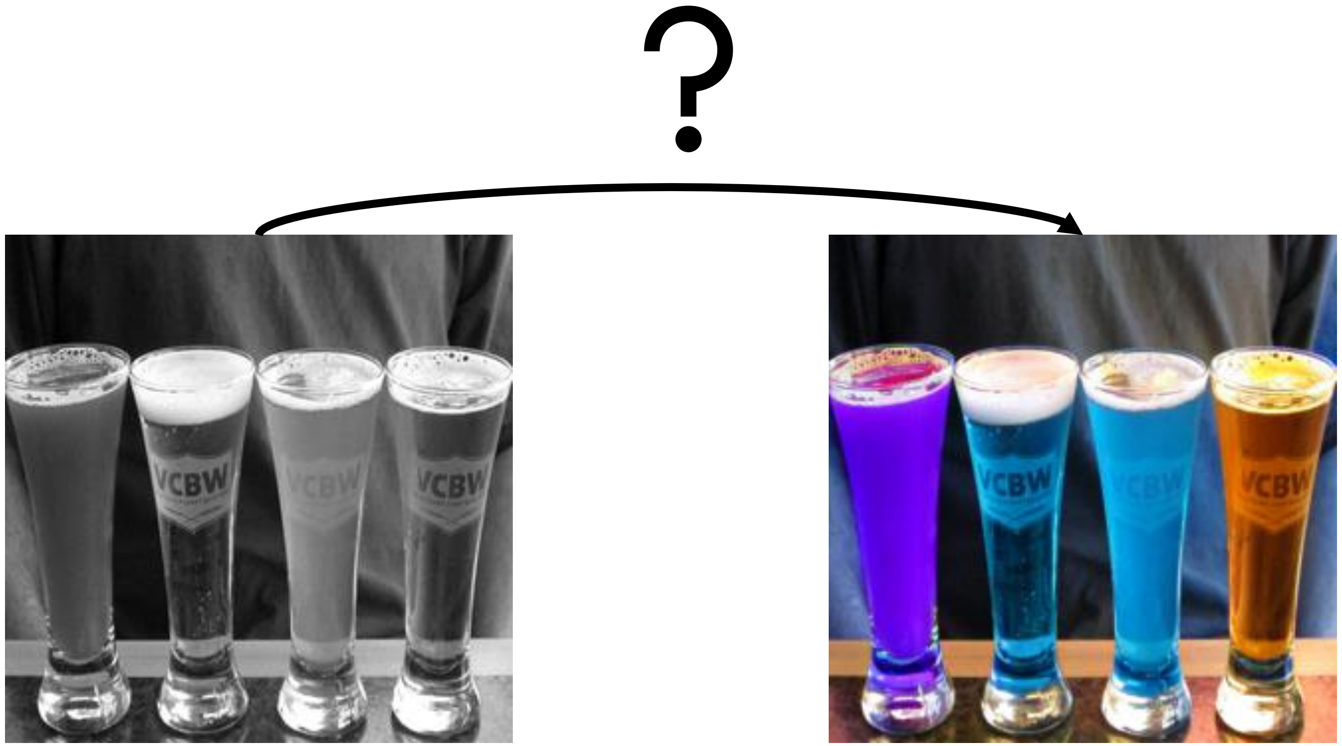
- Introduction
  - Background
  - Motivation
- Method
  - Pipeline
  - Luminance-guided image compression
  - Semantic-aligned latent representation
  - Instance-aware sampling strategy
- Result
  - Comparison with Automatic Colorization
  - Comparison with Language-based Colorization
  - Ablation
  - Application

Background

1. A person wearing a spacesuit is mowing grass on the lawn.
2. A rabbit with sunglasses and a hat.
3. An astronaut piled up a pyramid with sand.
4. The robot is pouring coffee……

Motivation

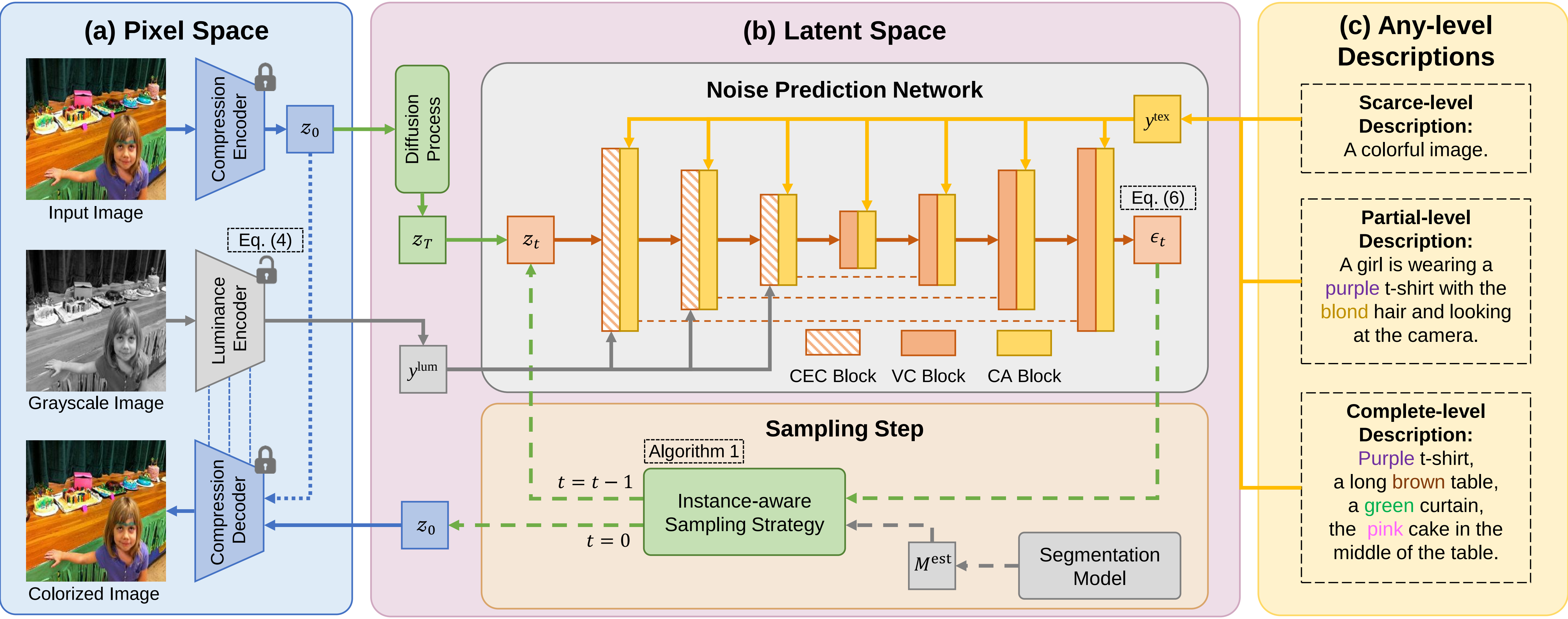
The wine on the far left is purple, the middle two glasses are blue, and the wine on the far right is orange.



Outline

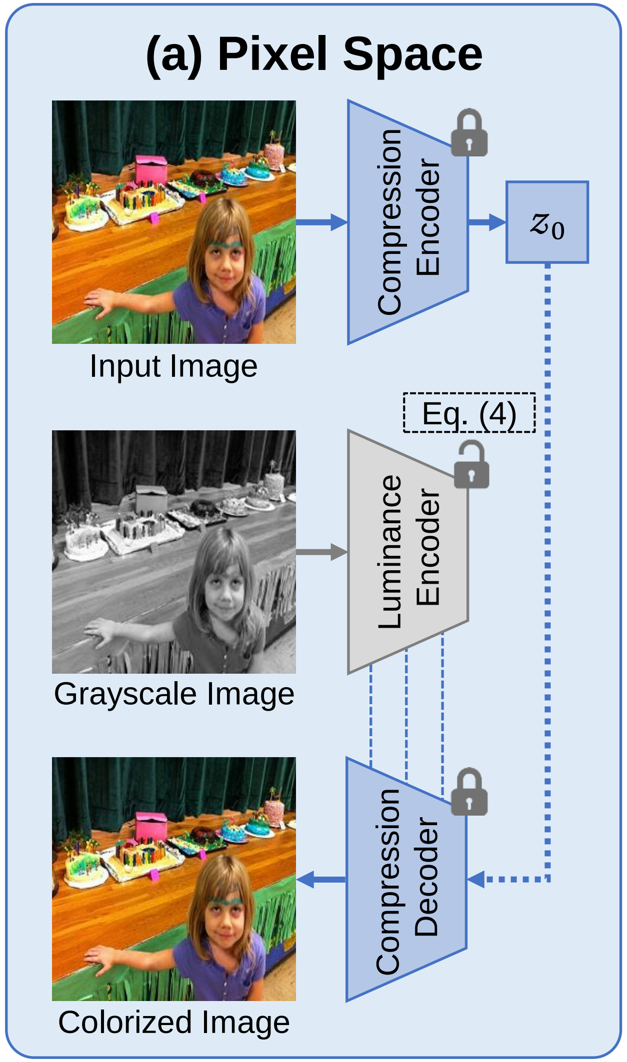
- Introduction
  - Background
  - Motivation
- Method
  - Pipeline
  - Luminance-guided image compression
  - Semantic-aligned latent representation
  - Instance-aware sampling strategy
- Result
  - Comparison with Automatic Colorization
  - Comparison with Language-based Colorization
  - Ablation
  - Application

Pipeline



(a) Pixel Space
Input Image
Compression Encoder
z_0
Eq. (4)
Grayscale Image
Luminance Encoder
Colorized Image
Compression Decoder
(b) Latent Space
Diffusion Process
z_T
Noise Prediction Network
z_t
y^{tex}
Eq. (6)
ϵ_t
y^{lum}
CEC Block
VC Block
CA Block
Sampling Step
Algorithm 1
$t = t - 1$
Instance-aware Sampling Strategy
$t = 0$
z_0
M^{est}
Segmentation Model
(c) Any-level Descriptions
Scarce-level Description:
A colorful image.
Partial-level Description:
A girl is wearing a purple t-shirt with the blond hair and looking at the camera.
Complete-level Description:
Purple t-shirt,
a long brown table,
a green curtain,
the pink cake in the middle of the table.

Luminance-guided image compression



Training loss:

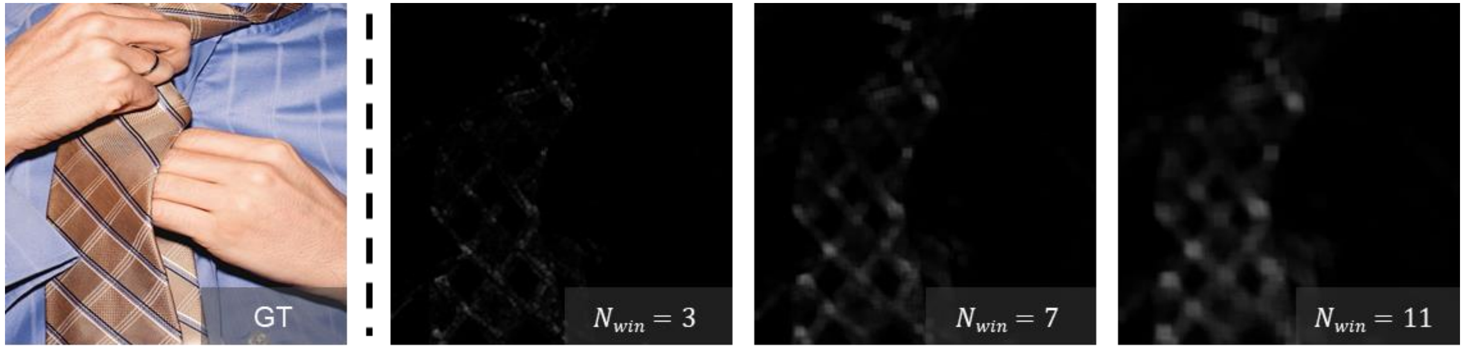
$$\mathcal{L}_{\text{pix}} = \mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{dis}} + \beta\mathcal{L}_{\text{per}}$$

New reconstruction loss:

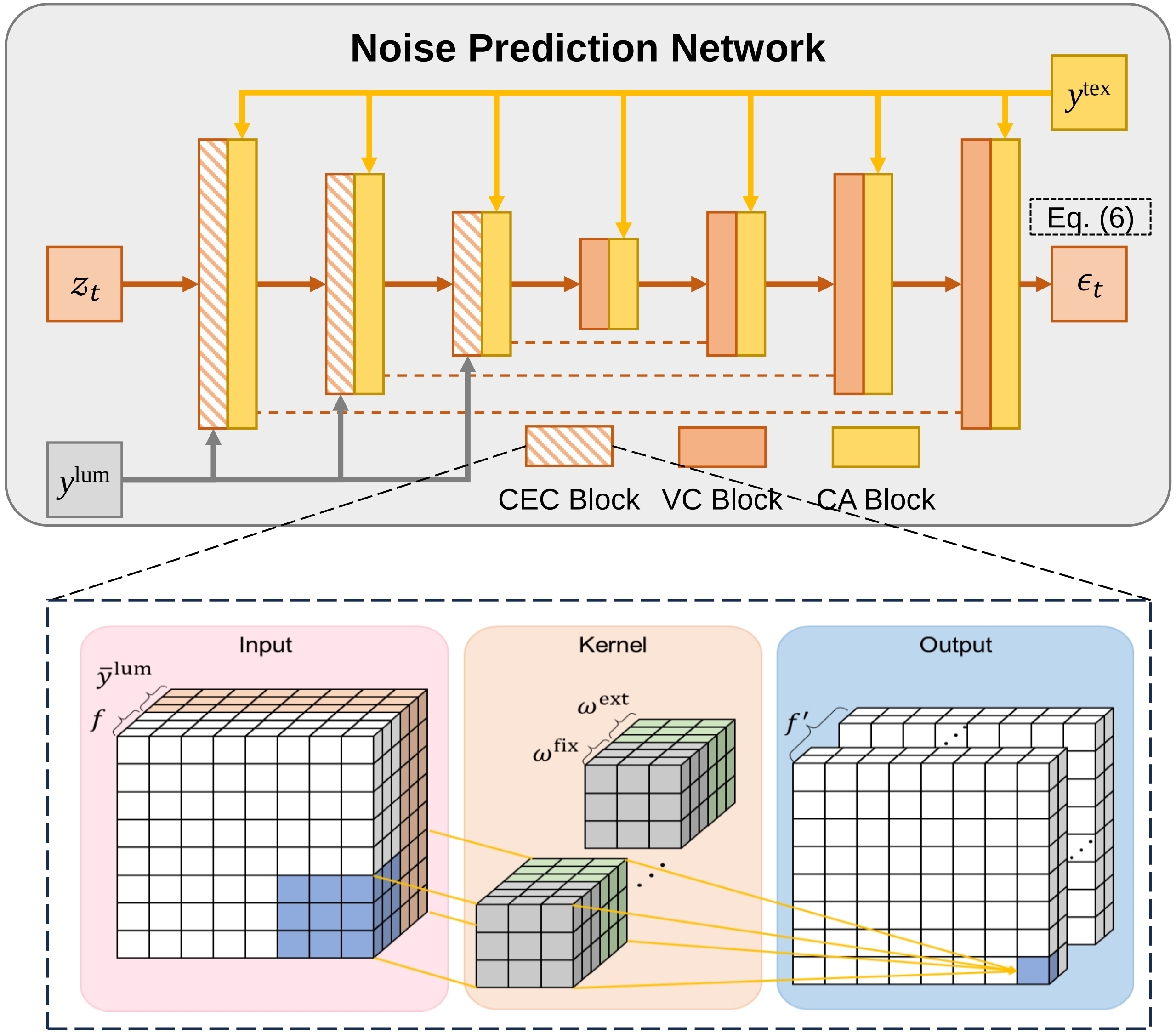
$$\mathcal{L}_{\text{rec}} = \| M^{\text{art}} \odot (x - \tilde{x}) \|_1.$$

Artifacts map:

$$M^{\text{art}}_{h,w} = \sum_{p \in \Omega_{(h,w)}} \left(\frac{\delta_p - \mu_p}{N_{\text{win}}} \right)^2, \qquad \mu_p = \sum_{p \in \Omega_{(h,w)}} \frac{\delta_p}{N^2_{\text{win}}},$$



Semantic-aligned latent representation



Formula for extended convolution:

$$f'_{h,w} = \sum_{i=0}^{N_{\text{k}}-1} \sum_{j=0}^{N_{\text{k}}-1} \Big(\sum_{k=1}^{N_{\text{fix}}} \omega^{\text{fix}}_{i,j,k} f_{p,q,k} + \sum_{k=1}^{N_{\text{ext}}} \omega^{\text{ext}}_{i,j,N_{\text{fix}}+k} \bar{y}^{\text{lum}}_{p,q,k} \Big),$$

Training loss:

$$\mathcal{L}_{\text{lat}} = \mathbb{E}_{t,z_0,\epsilon\sim\mathcal{N}(0,1)} \big[\| \epsilon_t - \epsilon_\theta(z_t, t, y^{\text{tex}}, y^{\text{lum}}) \|^2 \big],$$

Instance-aware sampling strategy

Algorithm 1: Instance-aware sampling strategy

input : Roughly estimated object contour M^{est}
output : Colorized latent representation z_0
for $t = T \ldots 1$ **do**
  $\_, M_*^{\text{att}} = \epsilon_\theta(z_t, t, y^{\text{lum}}, y^{\text{tex}})$
  for $l = 1 \cdots L$ **do**
    $\hat{M}_l^{\text{est}} \leftarrow \text{Downsampling}(M^{\text{est}}, l)$
    $\mathcal{M} \leftarrow \text{Sigmoid}(M_l^{\text{att}})$
    $\hat{M}_l^{\text{att}} \leftarrow M_l^{\text{att}} - \lambda \nabla_{M_l^{\text{att}}} \mathcal{L}_{\text{BCE}}(\mathcal{M}, \hat{M}_l^{\text{est}})$
  end
  $\hat{\epsilon}_t, \_ = \epsilon_\theta(z_t, t, y^{\text{lum}}, y^{\text{tex}})\{M_*^{\text{att}} \leftarrow \hat{M}_*^{\text{att}}\}$
  $z_{t-1} = \text{DDIM}(z_t, \hat{\epsilon}_t, t)$
end

Ground-truth

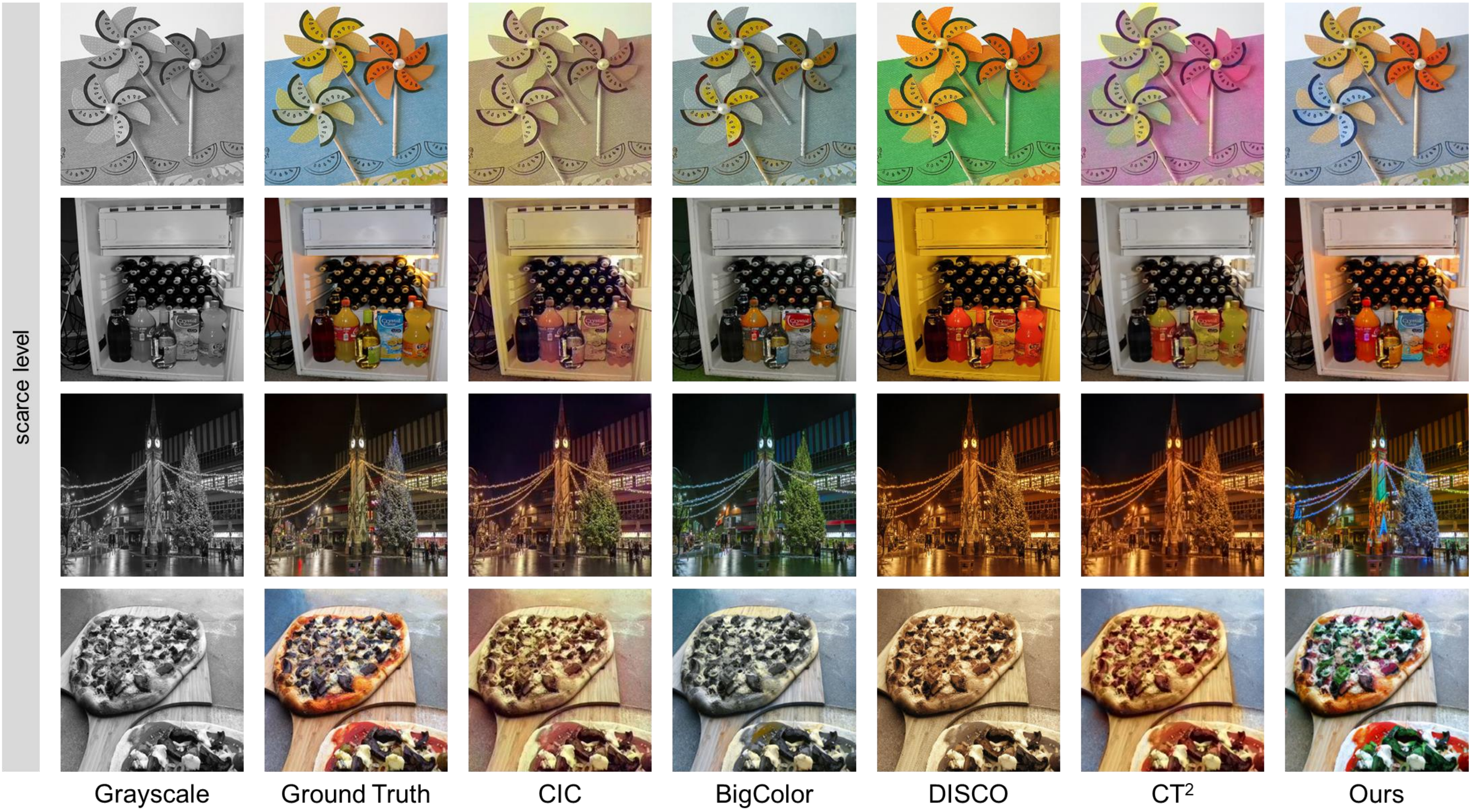
M^{est} for "left orange car"

M^{est} for "right yellow car"

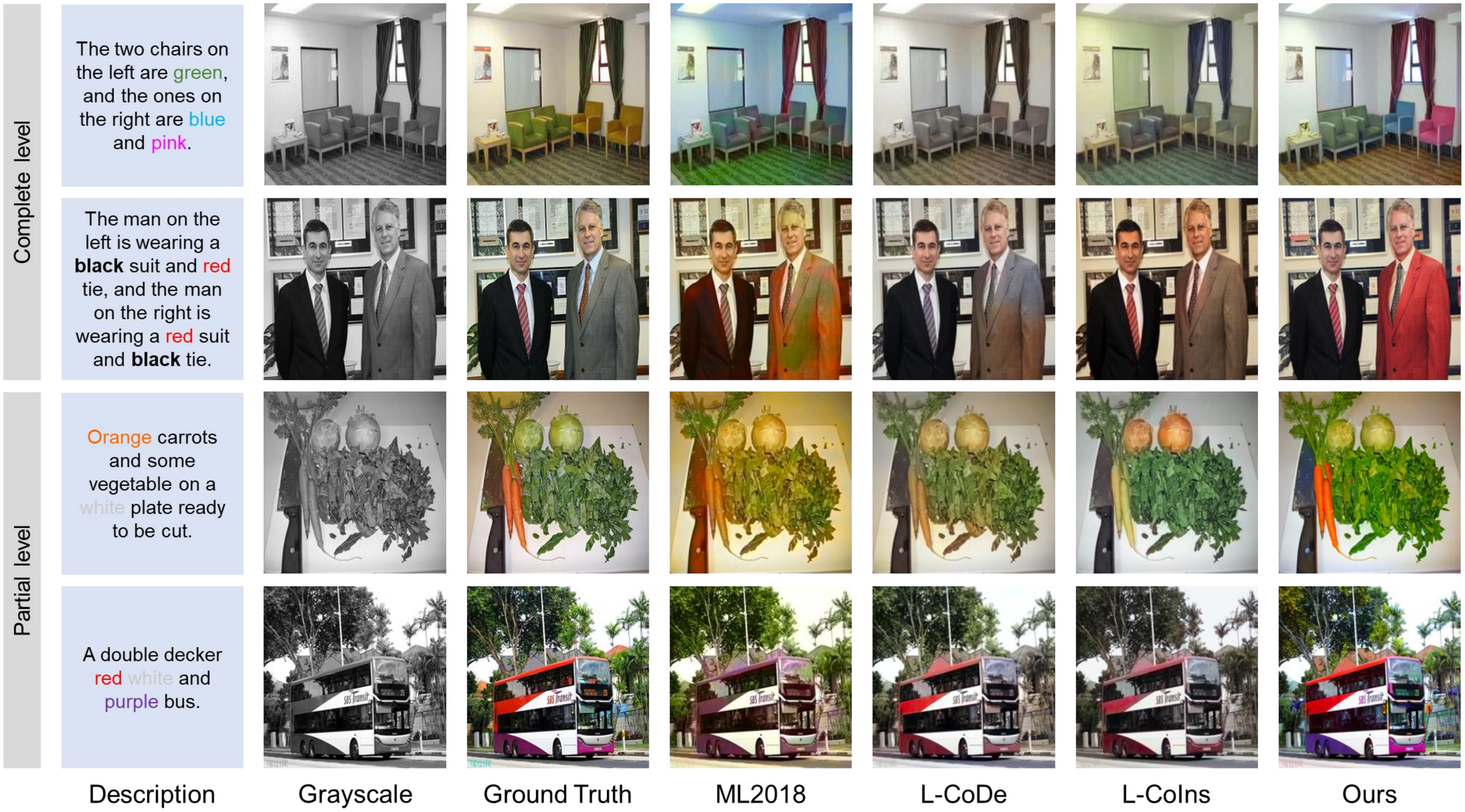
Outline

- Introduction
    - Background
    - Problems
- Method
    - Pipeline
    - Luminance-guided image compression
    - Semantic-aligned latent representation
    - Instance-aware sampling strategy
- Result
    - Comparison with Automatic Colorization
    - Comparison with Language-based Colorization
    - Ablation
    - Application

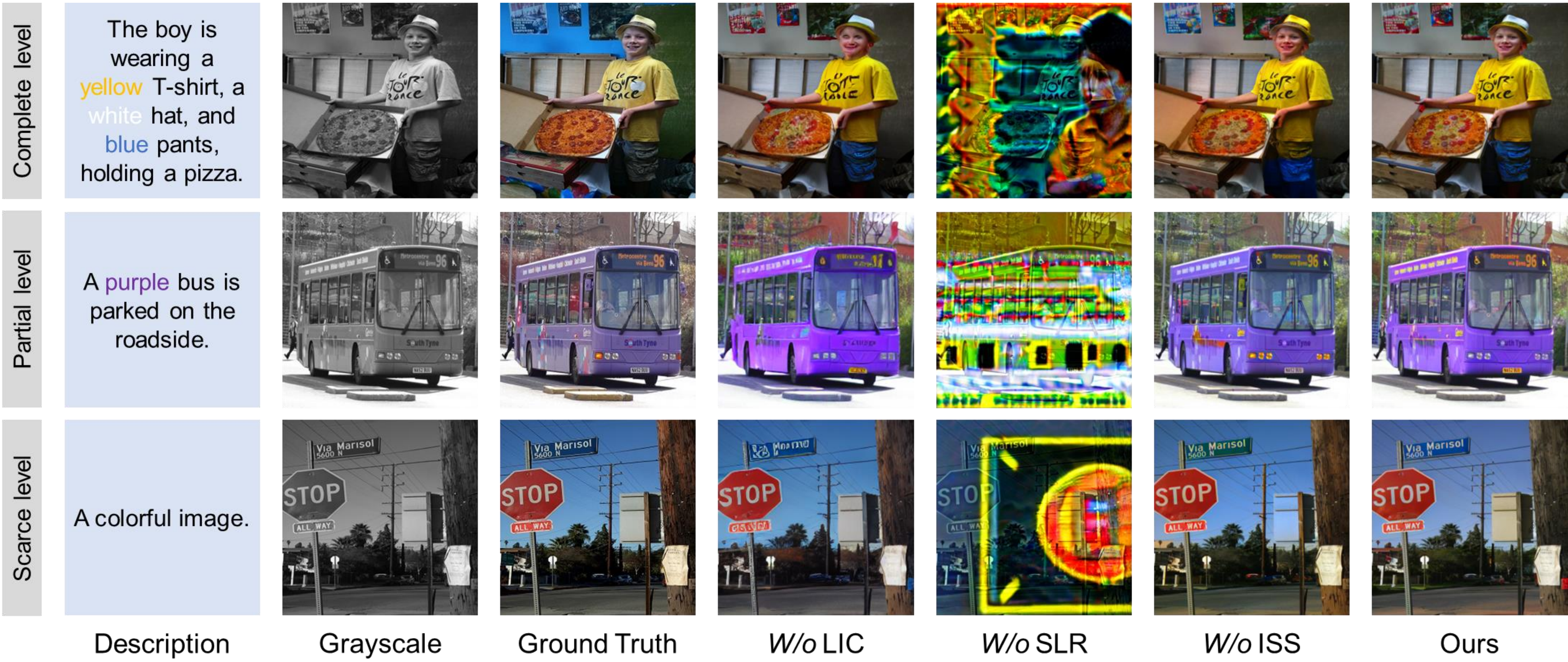
Comparison with Automatic Colorization



Comparison with Language-based Colorization



Ablation



Complete level
The boy is wearing a yellow T-shirt, a white hat, and blue pants, holding a pizza.
Partial level
A purple bus is parked on the roadside.
Scarce level
A colorful image.
Via Marisol
STOP
ALL WAY
Description
Grayscale
Ground Truth
W/o LIC
W/o SLR
W/o ISS
Ours

Application

1960. "Judge Myles Paige. Columbus, Georgia."

In front of the green vegetable stand stood a man in yellow clothes and green trousers.

1957. "Plymouth vs. Ford on the streets of Oakland circa."

There was a cyan car in the middle of the road.

There was a yellow car in the middle of the road.

1940. " The family of Mr. Timothy Levy Crouch at their Thanksgiving Day dinner "

A colorful image.

1897. "Palm walk on Lake Worth, Palm Beach."

There are many green trees along the beige path.

1925. "Washington, D.C. Judge Geo. H. MacDonald & Geo. G. Adams."

The man on the left was wearing an orange suit and the man on the right is wearing a pink suit.

The man on the left was wearing a gray suit and the man on the right is wearing a khaki suit.

1940. "Stephen A. Lynch Jr. residence, Sunset Island, Miami Beach."

A colorful image.