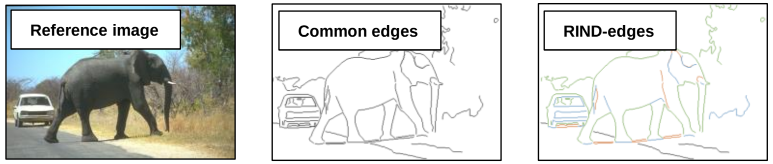
SWIN-RIND: Edge Detection for Reflectance, Illumination, Normal and Depth Discontinuity with Swin Transformer

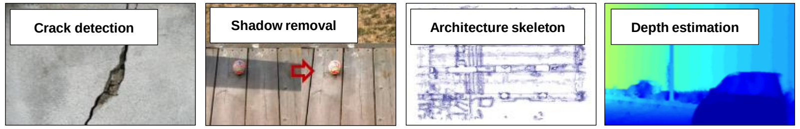
Miao Lun, Ishikawa Ryoichi, Oishi Takeshi
Institute of Industrial Science, The University of Tokyo

Introduction

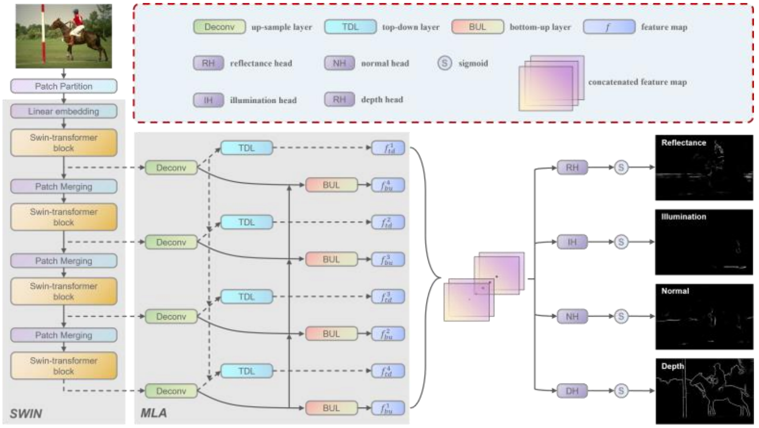
- Edges are caused by the discontinuities in Surface-Reflectance, Illumination, Surface-Normal, and Depth (RIND).
- Reflectance edges are caused by different textures.
- Illumination edges are caused by changes in light intensity.
- Normal edges appear at the intersection of planes.
- Depth edges are caused by mutation of depth.



- Application:



Proposed method



- The proposed network takes an encoder-decoder structure.
- Swin Transformer [2] extracts different levels of information.
- Multi-level feature aggregation(MLA) block integrates cues.
- Four independent decision heads predict RIND-edges simultaneously.

Loss function & optimization

- Self-balancing optimization using dice loss for fine edge detection
- Attention Loss [3]

$$L_a(Y,G) = -\sum_{(i,j)} \left(G_{(i,j)} \alpha \beta^{(1-Y_{(i,j)})^{\gamma}} \cdot \log(Y_{(i,j)})\right.$$
$$\left. + (1-G_{(i,j)})(1-\alpha)\beta^{Y_{(i,j)}^{\gamma}} \cdot \log(1-Y_{(i,j)})\right)$$

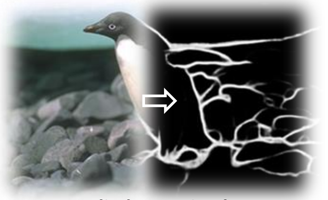
- Dice Loss [4]

$$L_d(Y,G) = 1 - \frac{2 \cdot \sum_{(i,j)} Y_{(i,j)} G_{(i,j)}}{\sum_{(i,j)} Y_{(i,j)}^2 + \sum_{(i,j)} G_{(i,j)}^2}$$

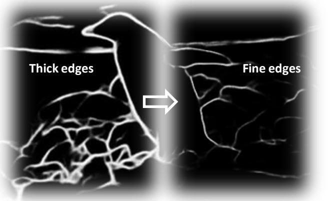
- Total Loss

$$L_t(Y,G) = \frac{1}{\rho^2} L_{a,r} + \frac{1}{\tau^2} L_{a,i} + \frac{1}{\varepsilon^2} L_{a,n} + \frac{1}{\mu^2} L_{a,d}$$
$$+ \eta \cdot \sum_k L_{d,k} + \log(\rho\tau\varepsilon\mu)$$

Y: *Predicted edge map*
G: *Ground truth edge map*



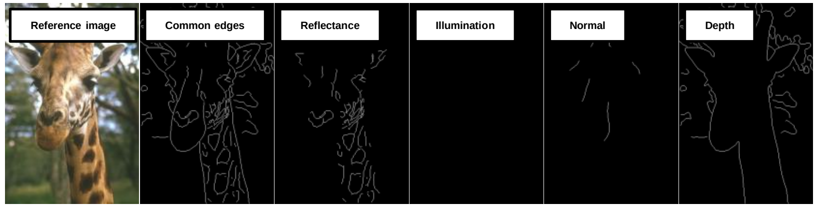
Applied attention loss



Dice loss and self-updating parameters

Dataset

- The proposed model was trained on the BSDS-RIND dataset [1].
- BSDS-RIND is an edge detection dataset which appends RIND-edge labels based on BSDS dataset.
- An example inside the BSDS-RIND:



Reference

[1] Pu et al. (2021). RINDNet: Edge Detection for Discontinuity in Reflectance, Illumination, Normal and Depth. *ICCV 2021*.
[2] Liu et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *ICCV 2021*.
[3] Wang et al. (2018). Doobnet: Deep object occlusion boundary detection from an image. *ACCV 2018*.
[4] Lee R Dice. (1945). Measures of the amount of ecologic association between species.

Experimental results

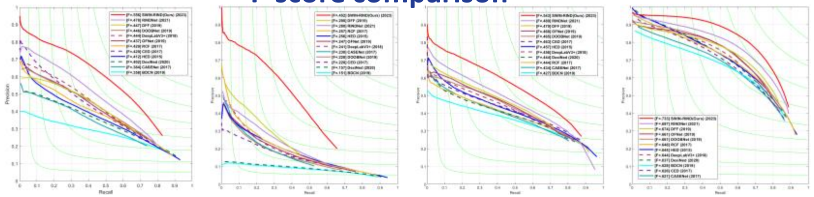
- From left to right: Reference image, Baseline, SWIN-RIND, Ground truth. Two tables below show the details of the ablation study.

MLA	BUL	TDL	ODS	OIS	AP
-	-	-	0.461	0.427	0.418
✓	×	✓	0.552	0.521	0.515
✓	✓	×	0.535	0.500	0.505
✓	✓	✓	**0.571**	**0.576**	**0.534**

L_a	L_d	SP1	SP2	CT	ODS	OIS	AP
✓	×	×	×	×	0.492	0.476	0.439
✓	×	×	✓	✓	0.113	0.114	0.052
✓	×	✓	×	✓	0.425	0.407	0.349
✓	✓	×	×	×	0.489	0.464	0.441
✓	✓	✓	×	✓	0.573	0.561	0.557
✓	✓	×	✓	✓	**0.571**	**0.576**	**0.534**

MLA: Multi-level feature aggregation, BUL: Bottom-up layer, TDL: Top-down layer
La: Attention loss, Ld: Dice loss, SP1: Self-learning parameter $1/\{\rho, \tau, \varepsilon, \mu\}$
SP2: Self-learning parameter $1/\{\rho^2, \tau^2, \varepsilon^2, \mu^2\}$, **CT:** Constraint term $\log(\rho\tau\varepsilon\mu)$

F-score comparison



Conclusion & limitation

- The proposed method SWIN-RIND outperforms state-of-the-art methods both in accuracy and visual effect.
- The experiment results show an adaptive combination of attention loss and dice loss is more effective in realizing fine edge detection.
- LIM: BSDS-RIND is the only available dataset for RIND-edge training.