Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Polymer Banknote Recognition and Tear Detection under Adverse Illumination: A Dataset and Local-Contrast Gated Attention

Quoc Thai Mai

23133071@student.hcmute.edu.vn

Ho Chi Minh City University of Technology and Engineering

Xuan Phi Nguyen

Ton Duc Thang University

Ngoc Tien Ho

Ho Chi Minh University of Banking

Huy Gia Toan Nguyen

Ho Chi Minh University of Banking

Research Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11662897/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Polymer Banknote Recognition and Tear Detection under Adverse Illumination: A Dataset and Local-Contrast Gated Attention

Quoc Thai Mai* · **Xuan Phi Nguyen*** · **Ngoc Tien Ho** · **Huy Gia Toan Nguyen**

Received: date / Accepted: date

Abstract Thin, low-contrast structures on glossy surfaces are easily lost when specular highlights saturate the camera sensor. We study this problem on polymer banknotes and introduce VN-Polymer-Adverse, a dataset of 1,812 images of Vietnamese notes (six denominations and tears) captured in six acquisition regimes, together with IR-YOLOv8n, a YOLOv8n variant with two Local-Contrast Gated Attention (LCGA) blocks in the neck. LCGA passes a local gradient-energy map to a signed spatial gate that can raise or lower neck activations. With 3.02 M parameters, IR-YOLOv8n reaches 98.76% mean average precision at an intersection-over-union threshold of 0.5 (mAP@50), against 97.24% for YOLOv8n, and raises tear average precision (AP@50–95) under glare from 62.47% to 67.96%. Contrast-limited adaptive histogram equalisation preprocessing gave no gain. IR-YOLOv8n is comparable to YOLOv8s on tears but trails it on denominations, and its latency on a graphics processing unit (GPU) is 42% higher than that of YOLOv8n. Each model was trained once, and the feature-level analysis did not confirm the mechanism that motivated the gate, so we present LCGA as a component that helps empirically.

Keywords Polymer banknotes · Tear detection · Specular glare · Attention mechanism · Lightweight object detection · YOLO

Quoc Thai Mai
Faculty of Information Technology, Ho Chi Minh City University of Technology and Engineering, Ho Chi Minh City, Vietnam E-mail: 23133071@student.hcmute.edu.vn

Xuan Phi Nguyen
Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam E-mail: nguyenxuanphi@tdtu.edu.vn

Ngoc Tien Ho
Faculty of Management Information System, Ho Chi Minh City University of Banking, Ho Chi Minh City, Vietnam E-mail: 030239230244@st.buh.edu.vn

Huy Gia Toan Nguyen
Faculty of Management Information System, Ho Chi Minh City University of Banking, Ho Chi Minh City, Vietnam E-mail: 030239230247@st.buh.edu.vn

1 Introduction

Camera-based inspection of glossy objects is difficult when specular highlights saturate the sensor and hide thin, low-contrast structure. Polymer banknotes are a practical case. Denomination recognition and fitness screening of banknotes are core functions of cash-processing machines and handheld validators [24, 13]. Many central banks have replaced cotton-paper notes with polymer notes printed on biaxially oriented polypropylene (BOPP), which carry see-through windows and other security features [32]. The smooth, glossy substrate that makes these notes durable also changes how they interact with light. Deep learning has advanced banknote analysis [24,4,25,22,28], but the studies surveyed in Sec. 2 neither vary illumination systematically for polymer notes nor annotate physical tears together with denominations.

Outside controlled enclosures, such as retail counters and outdoor kiosks, two lighting effects matter. Specular glare saturates pixels and washes out detail and colour [1]; once a pixel is clipped, its local gradient is lost. Back-lighting makes the clear window and translucent coating of a polymer note transmit background structure [32], which may cause spurious detections. Tears, which matter for fitness classification [13,11], are thin, low-contrast structures that glare can hide.

The usual remedy is to enhance the image before detection, using contrast-limited adaptive histogram equalisation (CLAHE) [36], Retinex-based decomposition [12], learned curve estimation [2] or image-adaptive filtering trained jointly with the detector [18]. These methods target under-exposure or haze, add a processing stage, and cannot recover detail that clipping has already removed. In our experiments, applying CLAHE before a YOLOv8n trained on unenhanced images gave no gain (96.97% vs. 97.24% mean average precision at an intersection-over-union (IoU) threshold of 0.5, abbreviated mAP@50) and added 3.20 ms per image.

We explore a detector-internal alternative. IR-YOLOv8n adds Local-Contrast Gated Attention (LCGA) to the neck of YOLOv8n, a feature pyramid network [15] with path aggregation [17] (PAN-FPN). LCGA passes a local gradient-energy map, $D(x) = \mathrm{Mean}_c(|x - \mathrm{AvgPool}_{3\times3}(x)|)$, to a spatial gate, on the premise that clipped regions have little local variation whereas tear edges and numerals have a lot. Section 6.6 examines whether this premise holds in neck features and what the trained gate does.

The contributions of this work are as follows:

1. **VN-Polymer-Adverse.** A dataset of polymer banknotes captured under adverse illumination, with tear annotations: 1,812 images and 2,249 boxes over seven classes (six denominations and tear) in six regimes (indoor reference, outdoor sunlight, backlight, glare, and tears under neutral light and under glare).
2. **IR-YOLOv8n.** A YOLOv8n variant with C2f-LCGA blocks at neck layers 15 ($P_3/8$) and 18 ($P_4/16$). Each block combines a channel gate with a signed spatial gate driven in part by $D(x)$, and its output is added to the C2f output through a learnable scale initialised at 0.05. We compare LCGA with five generic attention modules and ablate the gate design, the differential operator, the pooling window and the insertion level.
3. **Benchmark and cost analysis.** Against eight baselines, IR-YOLOv8n reaches 98.76% mAP@50 with 3.02 M parameters and raises average precision (AP@50–95) on tears under glare by 5.49 points over YOLOv8n, at a 42% higher latency on a T4 GPU. It is comparable to YOLOv8s on tears and behind it on denominations.

We evaluate the module on Vietnamese polymer notes only; whether it transfers to other glossy-surface inspection tasks is untested (Sec. 7.1).

2 Related Work

2.1 Banknote recognition and fitness inspection

Banknote analysis spans denomination recognition, counterfeit detection and fitness classification [13], and most work relies on dedicated sensing: line-sensor CNNs [24], fuzzy fitness assessment from visible-reflection and near-infrared transmission images [11], joint recognition and counterfeit detection [4], and fake-note detection in smartphone images [25]. For polymer notes, Wang et al. [32] use the random translucent pattern of the back-lit substrate as a fingerprint. In their work back-lighting is a controlled signal, whereas in ours it is an uncontrolled nuisance. Detectors have also been applied directly: YOLOv8 to Pakistani currency [22] and a YOLOv8n variant with squeeze-and-excitation blocks to US dollar, euro and taka notes [28].

Two public resources are closer to our setting. Meshram et al. [20] release 5,125 images of Indian paper banknotes, including soiled, torn and crumpled notes, organised by denomination; to our knowledge, the damage is not annotated with bounding boxes. BankNote-Net [23] collects 24,826 images of 112 denominations from 17 currencies under varied lighting and background, but its public release consists of image embeddings rather than annotated images. Among the works surveyed here, only one concerns polymer notes [32], and none, to our knowledge, localises tears or reports per-illumination results for polymer notes: Kwon et al. set crumpling and tears aside [11], and the detector studies [22, 28] give no per-condition breakdown.

2.2 Adverse illumination, lightweight detectors and attention

Enhancement before detection is the usual response to poor illumination. CLAHE [36] equalises local histograms under a contrast limit, Retinex theory [12] separates reflectance from illumination, Zero-DCE [2] learns light-enhancement curves, and IA-YOLO [18] predicts the parameters of a differentiable image-processing module jointly with a YOLO detector. These methods target under-exposure or haze rather than saturated highlights. Highlight-removal methods [1], including text-aware ones [6], run as a separate network or processing stage, and learned variants often need paired real-world data. Both families add a stage before the detector, whereas IR-YOLOv8n modulates neck features inside it.

The YOLO family [26] now includes YOLOv8 [9], YOLOv9 [30], YOLOv10 [29], which removes non-maximum suppression (NMS), and YOLO11 [10]; the Ultralytics package [9] also provides an anchor-free YOLOv5u variant, evaluated here as YOLOv5nu. GhostNet [3] reduces cost with cheap linear operations. Lightweight attention is often added to such detectors: squeeze-and-excitation (SE) [7] and efficient channel attention (ECA) [31] reweight channels, the convolutional block attention module (CBAM) [33] adds a spatial map computed from channel-wise mean and max pooling, coordinate attention (CA) [5] encodes position along two axes, and SimAM [34] derives a parameter-free weight from the deviation of each activation from its channel mean. All of these use sigmoid gates with weights in $(0, 1)$. LCGA differs in two respects: it feeds a local-neighbourhood deviation map to a learned spatial gate, and its tanh gate lies in $(-1, 1)$, so that, in the residual form of Eq. (5), it can raise as well as lower the activation at a location. We test the two choices separately in Sec. 6.4.

3 The VN-Polymer-Adverse dataset

VN-Polymer-Adverse contains 1,812 camera images of circulating Vietnamese polymer notes (BOPP substrate). They cover the six denominations in circulation (10,000 to 500,000 Vietnamese dong, VND; written 10k to 500k below) and notes with real tears. All images are original captures; none is synthetic or a duplicated copy.

The images fall into six acquisition regimes (Table 1): an indoor reference under diffuse light (about 400 lux); outdoor direct sunlight (above 25,000 lux) with hard cast shadows; back-light, where a source behind the note makes the clear window and translucent coating transmit light; glare, where a directional spotlight or flash produces specular highlights; and two tear subsets, photographed under diffuse light (neutral) or with a highlight over the damage (glare). The tear subsets therefore differ from the other four regimes in damage as well as in lighting. Each image carries YOLO-format boxes for seven classes (the six denominations and *tear*). Within each regime the images were split roughly 70/15/15 into 1,260 training, 276 validation and 276 test images, containing 2,249 boxes (1,568, 343 and 338, respectively). The split was made per physical note, so no note in the test images was seen during training or validation. The test split holds 276 denomination boxes (216 on intact notes, 60 on torn notes) and 62 tear boxes in 60 images.

Table 1 Composition of VN-Polymer-Adverse (number of images).

Regime	Train	Val	Test	Total
Indoor reference	270	60	60	390
Outdoor sunlight	270	60	60	390
Strong back-light	210	48	48	306
Specular glare	210	48	48	306
Tear, neutral light	150	30	30	210
Tear, glare	150	30	30	210
Images	1,260	276	276	1,812
Boxes	1,568	343	338	2,249

4 Method

4.1 Image formation under specular highlights

A glossy surface returns a diffuse (body) component plus a specular lobe around the mirror direction [27]. With I_d and I_s the corresponding sensor responses, the recorded value at pixel (u, v) in each colour channel is

$$I(u, v) = \min\big(I_{\max},\, I_d(u, v) + I_s(u, v)\big), \tag{1}$$

with $I_{\max} = 255$ for 8-bit images. Where I_s is large the pixel clips and the image becomes locally constant, so ∇I vanishes except along the border of the region. A highlight over a thin structure such as a tear therefore removes the local variation that would reveal it, which motivates a gate driven by local variation. This argument concerns pixels, and Sec. 6.6 tests whether it carries over to neck features.

4.2 IR-YOLOv8n

IR-YOLOv8n keeps the YOLOv8n backbone and head. In the neck, the C2f blocks at layers 15 ($P_3/8$, 64 channels) and 18 ($P_4/16$, 128 channels) are replaced by C2f-LCGA blocks (Fig. 1); layer 21 ($P_5/32$) is unchanged.

Local gradient energy. For a feature map $x \in \mathbb{R}^{C\times H\times W}$ we define

$$D(x) = \mathrm{Mean}_c\big(|x - \mathrm{AvgPool}_{3\times 3}(x)|\big) \in \mathbb{R}^{1\times H\times W}, \tag{2}$$

with stride 1, padding 1 and Mean_c the channel average. $D(x)$ is the mean absolute deviation of each activation from its 3×3 neighbourhood, a parameter-free measure of local contrast that is small where the map is flat and large near edges and strokes. We call it a gradient-energy map by analogy with the pixel-level gradient magnitude; strictly, it is a box-filter high-pass response. Unlike the spatial branch of CBAM [33], which can follow bright regions with or without structure, $D(x)$ responds to variation rather than level.

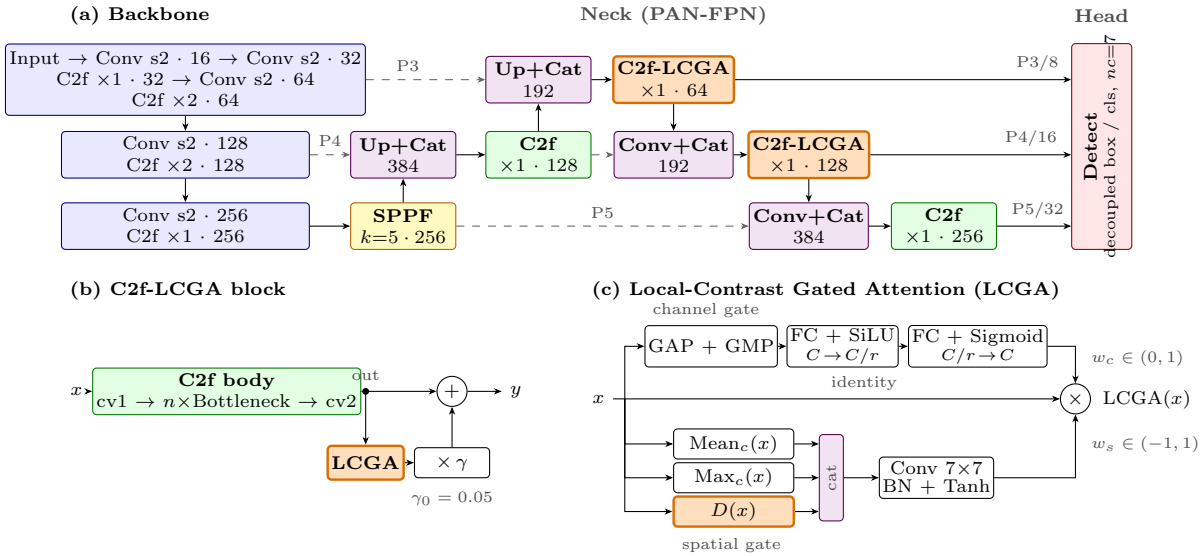


Fig. 1 IR-YOLOv8n. (a) YOLOv8n with C2f-LCGA blocks (orange) at neck layers 15 ($P_3/8$) and 18 ($P_4/16$); Conv is 3×3, s2 is stride 2, $\cdot c$ gives the number of output channels (nano width), and dashed arrows are skip links. (b) C2f-LCGA block, $y = \text{out} + \gamma\ \text{LCGA}(\text{out})$. (c) LCGA: channel gate $w_c \in (0, 1)$ with reduction ratio $r = 4$, spatial gate $w_s \in (-1, 1)$ computed from $\text{Mean}_c(x)$, $\text{Max}_c(x)$ and $D(x) = \text{Mean}_c|x - \text{AvgPool}_{3\times3}(x)|$; $\times$ is the element-wise product.

Local-contrast gated attention. The channel gate combines global average pooling (GAP) and global max pooling (GMP) in a bottleneck multilayer perceptron (MLP) with reduction ratio $r = 4$:

$$w_c = \sigma\big(W_2\ \text{SiLU}(W_1(\text{GAP}(x)+\text{GMP}(x)))\big) \in (0,1)^{C\times1\times1}, \tag{3}$$

where SiLU is the sigmoid-weighted linear unit. The spatial gate stacks the channel mean, the channel maximum and $D(x)$, and applies a 7×7 convolution, batch normalisation [8] and tanh:

$$w_s = \tanh\big(\text{BN}(\text{Conv}_{7\times7}([\text{Mean}_c(x);\ \text{Max}_c(x);\ D(x)]))\big) \in (-1,1)^{1\times H\times W}. \tag{4}$$

Then $\text{LCGA}(x) = x \odot (w_c \odot w_s)$. Unlike the sigmoid gates of SE, CBAM and similar modules, w_s can be negative. We initially hypothesised that $w_s < 0$ over flat, specular regions would mute glare blooming while preserving structure; Sec. 6.6 reports what training produced.

C2f-LCGA block. The gated term is added to the C2f output through a learnable scalar γ:

$$y = \text{out}+\gamma\ \text{LCGA}(\text{out}) = \text{out}\odot\big(1+\gamma\, w_c\odot w_s\big), \text{out} = \text{C2f}(x). \tag{5}$$

We set $\gamma_0 = 0.05$. Since $|w_c \odot w_s| < 1$, each activation changes by less than 5% at initialisation, so the block starts close to the original C2f. After training, $\gamma = 0.0358$ at layer 15 and 0.0743 at layer 18, so the modulation stays within 3.6% and 7.4% of the C2f output.

5 Experimental setup

5.1 Compared methods

We compare IR-YOLOv8n with six lightweight detectors (YOLOv8n [9], YOLO11n [10], YOLOv5nu, YOLOv10n [29], YOLOv9t [30] and Ghost-YOLOv8n [3]), with a pipeline in which CLAHE [36] enhances each image before an unmodified YOLOv8n trained on the original images, and with YOLOv8s as a larger reference. In the CLAHE pipeline, CLAHE is applied to the luminance channel in CIELAB space (clip limit 2.0, 8×8 tile grid) and the result is converted back to the original colour space, so the detector needs no retraining.

5.2 Training and evaluation protocol

All models were trained on the training split for 50 epochs at 640×640 with batch size 16 on one GPU, using Ultralytics (version 8.3 / PyTorch 2.9.0) with automatic mixed precision [21]. Training specified `optimizer: auto`, which resolves to AdamW ($\text{lr}_0 \approx 0.00091$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 5×10^{-4}) for runs under 10,000 iterations ($\approx$ 3,950 here), with a cosine annealing schedule decaying to $0.01 \times \text{lr}_0$ and 3 warmup epochs; weight decay is not

applied to normalisation and bias parameters. The detection loss is the standard Ultralytics combination of complete-IoU (CIoU) box regression [35] (weight 7.5), binary cross-entropy classification (weight 0.5) and distribution focal loss [14] (weight 1.5). Augmentation follows the Ultralytics defaults (mosaic, hue–saturation–value (HSV) jitter, random affine and scale, horizontal flip), with mosaic switched off for the last 10 epochs. Standard baselines were initialised from official COCO-pretrained [16] checkpoints. Ghost-YOLOv8n has no official Ultralytics release, so its GhostConv modules were trained from scratch (277/397 tensors transferred from YOLOv8n). IR-YOLOv8n transferred 319/373 tensors from YOLOv8n, and its LCGA residual scales were initialised at $\gamma_0 = 0.05$ so that training starts close to the pretrained C2f baseline. YOLOv8n itself transfers 319/355 tensors; the 36 tensors left out belong to the classification head, and IR-YOLOv8n leaves out the same 36 plus its 18 new LCGA tensors, while all C2f blocks at layers 15 and 18 are fully transferred.

The checkpoint with the highest validation mAP@50–95 was kept for test evaluation. All comparative models and ablation variants were trained under the same schedule and evaluated on the held-out test split (276 images), using a confidence threshold of 0.001 and an NMS IoU threshold of 0.65, so that all configurations are compared on the same basis.

5.3 Metrics and cost measurement

AP_c@50–95 is the average precision (AP) of class c averaged over IoU thresholds $0.50, 0.55, \ldots, 0.95$ [16]. We report mAP@50, the mean AP at IoU 0.5 over the seven classes, and mAP@50–95. Because the latter is an unweighted mean over classes, it satisfies $\mathrm{mAP@50–95} = \frac{1}{7}\left(6\,\mathrm{mAP}^{\mathrm{denom}}_{50–95} + \mathrm{AP}^{\mathrm{tear}}_{50–95}\right)$, where $\mathrm{mAP}^{\mathrm{denom}}_{50–95}$ is the mean over the six denominations. Because tear is only one class in seven, tear AP is always reported separately: on all 60 tear test images (overall, which is not the mean of the two subset values) and on each 30-image subset (glare tear AP is the value on the glare subset). The drop caused by glare is $\Delta = \mathrm{AP}^{\mathrm{tear}}_{50–95}(\text{neutral}) - \mathrm{AP}^{\mathrm{tear}}_{50–95}(\text{glare})$. Complexity is reported as the number of parameters and floating-point operations (FLOPs). Latency was measured on a Tesla T4 (batch size 1, FP32, 640×640, median of 50 timed passes after 10 warm-up passes, with GPU synchronisation around each timed pass and ablation variants measured in the same way); no embedded device was tested.

6 Results

6.1 Accuracy across regimes

Table 2 gives the comparison, and Table 3 lists precision, recall and the overall denomination and tear AP. IR-YOLOv8n reaches 98.76% mAP@50 and 93.14% mAP@50–95, which is 1.52 and 1.72 points above YOLOv8n (97.24% and 91.42%). It is ahead of the other lightweight detectors and of the CLAHE pipeline, which did not help (96.97% mAP@50, with 3.20 ms of extra time per image). YOLOv8s, which has 3.7 times the parameters and 3.5 times the FLOPs, is higher on overall denomination mAP@50–95 (97.65% vs. 96.83%) and on overall mAP@50–95 (93.75% vs. 93.14%), whereas IR-YOLOv8n is higher on mAP@50 (98.76% vs. 98.57%) and on overall tear AP@50–95 (70.99% vs. 70.32%). The differences are between 0.2 and 0.8 points, which a single run cannot resolve (Sec. 7.1), so we do not rank the two models.

On the intact-note regimes, the gain over YOLOv8n grows with the difficulty of the lighting: 0.09 points indoors, 1.06 outdoors, 1.99 under back-light and 2.07 under glare. From indoor to glare, denomination mAP@50–95 falls by 1.23 points for IR-YOLOv8n and 1.28 for YOLOv8s, but by 3.21 for YOLOv8n and by 4.0 to 4.4 for the other lightweight detectors and the CLAHE pipeline. YOLOv8s stays 0.4 to 0.9 points above IR-YOLOv8n in all four regimes (per-regime profile: Fig. S3). Weighted by image count over the four regimes (216 images), denomination mAP@50–95 is 97.36% for IR-YOLOv8n, 98.07% for YOLOv8s and 96.13% for YOLOv8n.

6.2 Tear detection under glare

YOLOv8n reaches 71.37% tear AP@50–95 under neutral light and 62.47% under glare ($\Delta = 8.90$); IR-YOLOv8n reaches 74.39% and 67.96% ($\Delta = 6.43$), a glare gain of 5.49 points (Fig. 2). With the CLAHE pipeline, tear AP falls from 71.04% to 61.68%, so enhancement did not protect the tears. YOLOv10n has the largest drop (10.49). The small drop of Ghost-YOLOv8n (6.51) reflects its low neutral AP (50.60%) rather than robustness, and probably stems from training its GhostConv modules from scratch without official pretraining. YOLOv8s scores 73.54% and 67.69% ($\Delta = 5.85$): IR-YOLOv8n is 0.85 and 0.27 points higher on the two subsets, yet YOLOv8s has the smaller drop. At this sample size a single missed or shifted box moves AP noticeably (Sec. 7.1), so we regard the two as comparable.

Table 2 Results on the VN-Polymer-Adverse test split. Denomination columns: mAP@50–95 (%) on the four intact-note regimes (60, 60, 48 and 48 images). Tear columns: AP@50–95 (%) on the two 30-image tear subsets; Δ is neutral minus glare. Bold: best in column; underline: best among models with at most 3.02 M parameters. Ghost-YOLOv8n has no official pretrained weights and was only partly initialised from YOLOv8n (Sec. 5).

	Params	FLOPs	Lat.	mAP	mAP	Denomination				Tear		
Model	(M)	(G)	(ms)	@50	@50–95	Indoor	Outdoor	Back-light	Glare	Neutral	Glare	$\Delta\downarrow$
IR-YOLOv8n (ours)	3.02	8.10	7.24	**98.76**	93.14	97.67	97.88	97.22	96.44	**74.39**	**67.96**	6.43
YOLOv8s	11.13	28.45	9.65	98.57	**93.75**	**98.59**	**98.52**	**97.61**	**97.31**	73.54	67.69	**5.85**
YOLOv8n	3.01	8.09	**5.09**	97.24	91.42	97.58	96.82	95.23	94.37	71.37	62.47	8.90
CLAHE + YOLOv8n	3.01	8.09	8.29	96.97	90.98	97.47	96.43	94.82	93.48	71.04	61.68	9.36
YOLO11n	2.58	**6.32**	7.17	96.80	90.66	97.32	96.08	94.49	93.31	70.07	61.21	8.86
YOLOv5nu	2.50	7.07	6.99	96.53	90.06	96.97	95.73	94.02	92.78	68.17	59.73	8.44
YOLOv10n	2.27	6.53	6.31	96.40	89.81	96.68	95.42	93.73	92.37	69.45	58.96	10.49
YOLOv9t	**1.97**	7.61	14.08	95.99	88.97	96.43	94.97	93.28	92.02	65.36	56.48	8.88
Ghost-YOLOv8n	2.73	7.54	6.08	93.86	85.97	95.17	93.93	91.82	91.18	50.60	44.09	6.51

Table 3 Precision, recall and overall AP@50–95 (%) on the test split. Denom.: mean over the six denomination classes; Tear: tear class over all 60 tear test images. Together, Denom. and Tear give the mAP@50–95 of Table 2 as (6 Denom. + Tear)/7.

Model	Prec.	Rec.	Denom.	Tear
IR-YOLOv8n (ours)	97.88	98.86	96.83	70.99
YOLOv8s	98.67	98.36	97.65	70.32
YOLOv8n	96.73	96.44	95.54	66.72
CLAHE + YOLOv8n	96.13	95.85	95.12	66.14
YOLO11n	96.41	95.55	94.86	65.43
YOLOv5nu	95.82	95.25	94.45	63.72
YOLOv10n	95.21	94.36	94.12	63.96
YOLOv9t	95.77	94.07	93.68	60.68
Ghost-YOLOv8n	93.58	90.80	92.45	47.07

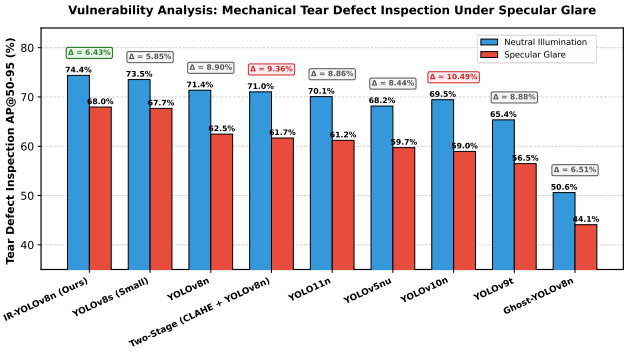


Fig. 2 Tear AP@50–95 under neutral light and glare for the compared models. "Two-Stage" denotes the CLAHE + YOLOv8n pipeline.

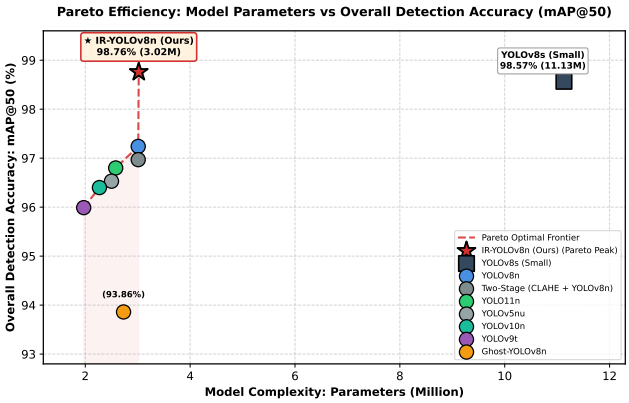


Fig. 3 Parameters versus overall mAP@50 for the compared models. "Two-Stage" denotes the CLAHE + YOLOv8n pipeline.

6.3 Computational cost

IR-YOLOv8n is almost as small as YOLOv8n (3.02 M vs. 3.01 M parameters, 8.10 vs. 8.09 GFLOPs; Fig. 3) but takes 7.24 ms (138 frames per second, FPS) instead of 5.09 ms (196 FPS), which is 42% more, so FLOPs understate its cost. At a neck level with C channels (and $r = 4$), LCGA adds $C^2/2 + 150$ parameters (two bias-free MLP layers, the 147 weights of the 7×7 convolution, two batch-normalisation parameters and γ): 2,198 at P_3 and 8,342 at P_4, 10,540 in total. The LCGA operations add little arithmetic and are probably limited by memory access [19], though we did not profile this. IR-YOLOv8n is slower than YOLO11n, YOLOv5nu, YOLOv10n and Ghost-YOLOv8n (6.08 to 7.17 ms) and faster only than YOLOv8s, YOLOv9t and the CLAHE pipeline (the last by 1.05 ms, with 1.79 points higher mAP@50).

6.4 Ablation study

Table 4 lists the ablations; parameters, FLOPs and the denomination metrics of the generic modules are in the supplementary material (Table S1).

Generic attention. SE, ECA, CBAM, CA and SimAM at the same two positions raise glare tear AP only from 62.47% to between 62.79% and 63.85%, and mAP@50 to between 97.35% and 97.68%, against 67.96% and 98.76% for LCGA. The five generic modules are within 1.1 points of each other, and the data do not allow us to separate them. They and LCGA all have 3.01–3.02 M parameters (Table S1), so the difference is not explained by model size. On the stricter

Table 4 Ablation analysis on the test split (276 images; mAP@50 in %, tear AP@50–95 in %). (a) is the CBAM row, used as the sigmoid-gated reference without $D(x)$; (d) is the adopted LCGA ($D(x)$ with $k = 3$, inserted at P_3 and P_4). Each row of the last three groups changes one design choice of (d): the operator that replaces $D(x)$, the pooling window k, or the neck level(s) at which LCGA is inserted.

		mAP	Tear AP		
Group	Variant	@50	Neutral	Glare	$\Delta\downarrow$
	YOLOv8n (no attention)	97.24	71.37	62.47	8.90
Generic modules	+ SE	97.38	71.55	62.84	8.71
	+ ECA	97.35	71.48	62.79	8.69
	+ CBAM (a)	97.45	71.85	63.12	8.73
	+ CA	97.68	72.25	63.85	8.40
	+ SimAM	97.60	72.10	63.70	8.40
Gate design	(b) tanh, no $D(x)$	97.92	72.64	64.85	7.79
	(c) sigmoid, with $D(x)$	98.15	73.10	65.48	7.62
	(d) tanh, with $D(x)$ (LCGA)	98.76	74.39	67.96	6.43
Operator	Sobel 3×3	98.24	73.40	66.21	7.19
	Laplacian 3×3	98.10	73.15	65.85	7.30
Window	$k = 5$	98.42	73.85	67.12	6.73
	$k = 7$	98.18	73.45	66.45	7.00
Insertion	P_3 only	98.15	73.65	66.18	7.47
	P_4 only	98.02	72.90	65.45	7.45
	P_5 only	97.48	71.90	63.20	8.70
	$P_3 + P_4 + P_5$	98.78	74.42	68.02	6.40

overall mAP@50–95 the generic modules reach 91.51% to 91.95%, against 93.14% for LCGA.

Gate design. Relative to (a), using tanh instead of a sigmoid (b) raises glare tear AP by 1.73 points, adding $D(x)$ to the sigmoid gate (c) raises it by 2.36, and using both (d) raises it by 4.84, which is 0.75 more than the sum of the two separate gains (4.09). Both ingredients seem to help, but the excess over additivity is too small to establish an interaction at this sample size, and since CBAM also differs from LCGA in its channel branch, the decomposition is approximate. Replacing $D(x)$ by a Sobel or Laplacian operator gives 66.21% and 65.85% glare tear AP, and larger pooling windows ($k = 5$ and 7) give 67.12% and 66.45%, all below the 67.96% obtained with $D(x)$ and $k = 3$.

Insertion. LCGA at P_3 or P_4 alone gives 66.18% and 65.45%, and at P_5 alone 63.20%, close to the baseline. Adding P_5 to $P_3 + P_4$ changes glare tear AP by 0.06 points but adds 32,918 parameters and raises latency from 7.24 to 8.05 ms (11%), so we use $P_3 + P_4$.

6.5 Qualitative examples

Figure 4 shows three illustrative test cases (confidence threshold 0.25), which are not statistical evidence. In Case 1 (100k, tear under glare), YOLOv8n misses the tear whereas IR-YOLOv8n finds it (confidence 0.52). In Case 2 (20k, tear under glare), both models find it, with confidences of 0.68 (YOLOv8n) and 0.90 (IR-YOLOv8n). In Case 3 (500k, back-light), YOLOv8n adds a spurious "10" box (confidence 0.46) that IR-YOLOv8n does not.

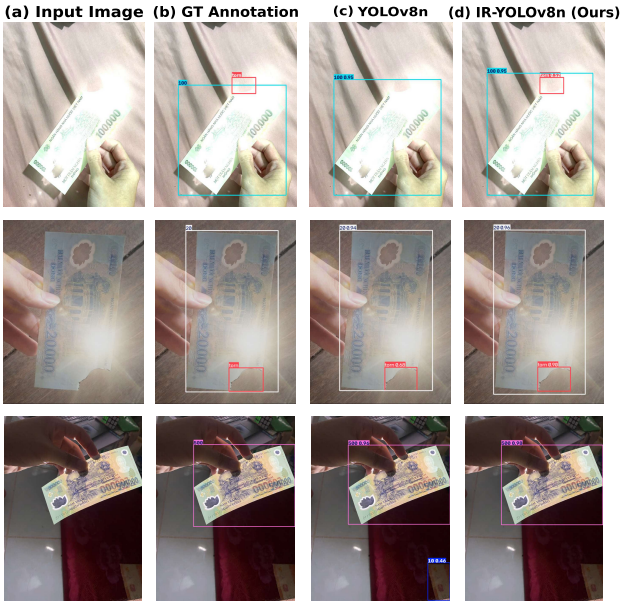


Fig. 4 Qualitative comparison: (a) input, (b) ground truth, (c) YOLOv8n, (d) IR-YOLOv8n. Case 1: tear under glare. Case 2: tear under glare. Case 3: back-lit 500k note.

Table 5 Mean ± standard deviation of $D(x)$ and of the gate w_s by region on the test split, and the share of cells with a negative gate. Cell counts are given in the text. Highlight and tear-box cells overlap where glare falls on a tear and are counted in both; the remaining region excludes both.

Layer	Region	$D(x)$	w_s	% $w_s < 0$
15 (P_3/8)	Highlight	0.0993 ± 0.063	-0.0778 ± 0.487	56.0
	Remaining	0.1111 ± 0.059	-0.0003 ± 0.503	47.2
	Tear box	0.1474 ± 0.093	-0.3021 ± 0.621	65.6
18 (P_4/16)	Highlight	0.0915 ± 0.050	-0.1372 ± 0.554	59.3
	Remaining	0.0954 ± 0.047	-0.0185 ± 0.566	49.2
	Tear box	0.1865 ± 0.091	-0.6703 ± 0.394	91.6

6.6 Does the gate behave as intended?

The case for $D(x)$ rests on the premise that highlights are flat. We checked this on the neck features of the trained IR-YOLOv8n, recording layers 15 and 18 on the 276 test images and grouping feature cells into highlight cells (input HSV value $V \geq 225$ as a brightness proxy; 194,303 and 48,623 cells), cells inside ground-truth tear boxes (7,610 and 2,032) and the remaining banknote foreground (1,526,470 and 382,661). Table 5 gives the statistics; precise cell definitions, an example activation map and a summary plot are in the supplementary material (Figs. S1–S2).

$D(x)$ separates tear boxes from highlights: 0.147 vs. 0.099 at layer 15 (about 1.5 times) and 0.187 vs. 0.092 at layer 18 (about 2.0 times). Highlights, however, are only slightly flatter than the remaining cells (0.099 vs. 0.111 and 0.092 vs. 0.095), with standard deviations of 0.05 to 0.06, so the premise holds only weakly in deep fea-

tures. The gate w_s is negative on average over highlights (-0.078 and -0.137; 56.0% and 59.3% of cells negative) and close to zero over the remaining cells (-0.0003 and -0.019). This is the hypothesised direction, but the shift is small compared with the spread of w_s (standard deviations of 0.49 to 0.57). Inside tear boxes w_s is more negative still (-0.302 and -0.670; 65.6% and 91.6% negative), so on average the gate attenuates tear regions instead of amplifying them. Tear boxes contain much intact substrate and w_s is scaled by $\gamma \leq 0.074$, so these statistics do not tell us how the gate contributes to the gain. Cells are spatially correlated, so we report these figures descriptively and attach no significance tests to them.

7 Discussion

Two observations frame the results. First, the benefit of LCGA grows with the difficulty of the lighting (Sec. 6), and the five generic attention modules at the same neck positions recover at most 1.4 of the 5.5 points gained on glare tears (Table 4). Second, CLAHE did not help: once highlights clip, as in Eq. (1), a preprocessing stage cannot restore the lost variation. Both are consistent with a benefit from conditioning the neck on local contrast, but neither shows it.

The feature-level analysis (Sec. 6.6) does not confirm the hypothesised mechanism, so we treat $D(x)$ and the signed gate as components that help empirically and not as an explanation. Settling the question would need controls that the present study did not run: replacing $D(x)$ with a randomised map of matching statistics and freezing the gate at its initial value.

On cost, IR-YOLOv8n matches YOLOv8s on tears with 27% of its parameters and is still faster than YOLOv8s on the T4 (7.24 vs. 9.65 ms), but its latency is 42% higher than that of YOLOv8n. Whether this trade is worthwhile depends on the target hardware (Sec. 7.1).

7.1 Limitations

This study has several limitations. First, the benchmark covers Vietnamese polymer notes captured with a handheld camera; transfer to other polymer currencies with different transparent-window designs (for example Australian or Canadian dollars) was not tested. Second, the defect annotations cover tears only; other forms of damage, such as deep creases, substrate warping and window delamination, are not annotated. Third, latency was measured only on a Tesla T4 GPU, so whether the small parameter count and FLOPs carry over to embedded hardware, including after quantisation, is untested. Finally, each model was trained once and evaluated on 276 test images, with about 30 tear instances per condition, so small differences between models should not be over-interpreted.

8 Conclusion

We introduced VN-Polymer-Adverse, 1,812 images of Vietnamese polymer notes in six acquisition regimes with denomination and tear annotations, and IR-YOLOv8n, a YOLOv8n variant with two Local-Contrast Gated Attention blocks in the neck. On the test split, IR-YOLOv8n raises mAP@50 from 97.24% to 98.76% and tear AP@50–95 under glare from 62.47% to 67.96% with 3.02 M parameters, at a 42% higher latency than YOLOv8n on a T4 GPU. CLAHE preprocessing gave no gain, and five generic attention modules recovered at most 1.4 of the 5.5 points gained on glare tears. IR-YOLOv8n is comparable to YOLOv8s on tears and trails it on denominations. The feature-level analysis did not confirm the mechanism that motivated $D(x)$, so we present LCGA as a component that helps empirically, not as an explained one. Natural next steps are repeated runs with different seeds, the controls described in Sec. 7 (a randomised $D(x)$ map and a frozen gate), other polymer currencies and embedded hardware.

Declarations

Funding. The authors received no specific funding for this work.

Acknowledgements. The authors thank Ho Chi Minh City University of Technology and Engineering (HCMUTE) for providing the computational resources used in this research, and the Vietnam Bank for Agriculture and Rural Development (Agribank) for providing the banknote samples used in this study.

Competing interests. 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.

CRediT authorship contribution statement. **Quoc Thai Mai**: Conceptualization, Methodology, Resources, Software, Investigation, Formal analysis, Data curation, Writing – original draft, Visualization, Project administration, Validation. **Xuan Phi Nguyen**: Conceptualization, Methodology, Supervision, Writing – review & editing. **Ngoc Tien Ho**:

Resources, Data curation, Validation. **Huy Gia Toan Nguyen**: Software, Data curation, Validation.

Data and code availability. The source code is publicly available at `https://github.com/tmai110605/IR-YOLOv8n`. The dataset can be made available by the corresponding author on reasonable request, subject to applicable Vietnamese regulations and any required authorisation.

References

1. Artusi, A., Banterle, F., Chetverikov, D.: A survey of specularity removal methods. Computer Graphics Forum **30**(8), 2208–2230 (2011). DOI 10.1111/j.1467-8659.2011.01971.x
2. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1780–1789 (2020)
3. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: GhostNet: More features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1580–1589 (2020)
4. Han, M., Kim, J.: Joint banknote recognition and counterfeit detection using explainable artificial intelligence. Sensors **19**(16), 3607 (2019). DOI 10.3390/s19163607
5. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13713–13722 (2021)
6. Hou, S., Wang, C., Quan, W., Jiang, J., Yan, D.M.: Text-aware single image specular highlight removal. In: Pattern Recognition and Computer Vision (PRCV 2021), *Lecture Notes in Computer Science*, vol. 13022, pp. 115–127 (2021). DOI 10.1007/978-3-030-88013-2_10
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018)
8. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. Int. Conf. Machine Learning (ICML), *Proceedings of Machine Learning Research*, vol. 37, pp. 448–456 (2015)
9. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8. Software, version 8.0.0 (2023). URL `https://github.com/ultralytics/ultralytics`
10. Jocher, G., Qiu, J.: Ultralytics YOLO11. Software, version 11.0.0 (2024). URL `https://github.com/ultralytics/ultralytics`
11. Kwon, S.Y., Pham, T.D., Park, K.R., Jeong, D.S., Yoon, S.: Recognition of banknote fitness based on a fuzzy system using visible light reflection and near-infrared light transmission images. Sensors **16**(6), 863 (2016). DOI 10.3390/s16060863
12. Land, E.H.: The Retinex theory of color vision. Scientific American **237**(6), 108–128 (1977)
13. Lee, J.W., Hong, H.G., Kim, K.W., Park, K.R.: A survey on banknote recognition methods by various sensors. Sensors **17**(2), 313 (2017). DOI 10.3390/s17020313
14. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33 (2020)
15. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 2117–2125 (2017)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proc. European Conf. Computer Vision (ECCV), *Lecture Notes in Computer Science*, vol. 8693, pp. 740–755 (2014)
17. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 8759–8768 (2018)
18. Liu, W., Ren, G., Yu, R., Guo, S., Zhu, J., Zhang, L.: Image-adaptive YOLO for object detection in adverse weather conditions. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 1792–1800 (2022). DOI 10.1609/aaai.v36i2.20072
19. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In: Proc. European Conf. Computer Vision (ECCV), *Lecture Notes in Computer Science*, vol. 11218, pp. 122–138 (2018). DOI 10.1007/978-3-030-01264-9_8
20. Meshram, V., Meshram, V., Patil, K., Suryawanshi, Y., Chumchu, P.: A comprehensive dataset of damaged banknotes in Indian currency (Rupees) for analysis and classification. Data in Brief **51**, 109699 (2023). DOI 10.1016/j.dib.2023.109699
21. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H.: Mixed precision training. In: Int. Conf. Learning Representations (ICLR) (2018)
22. Mushtaq, A., Khan, H.W., Salman, M., Abid, F., Rasheed, J., Alsubai, S.: Helping blind individuals: Pakistani currency recognition system for blind people using the YOLOv8 model. U.P.B. Scientific Bulletin, Series C **87**(4), 139–156 (2025)
23. Oviedo, F., Vinnakota, S., Seleznev, E., Malhotra, H., Shaikh, S., Lavista Ferres, J.: BankNote-Net: Open dataset for assistive universal currency recognition (2022)
24. Pham, T.D., Lee, D.E., Park, K.R.: Multi-national banknote classification based on visible-light line sensor and convolutional neural network. Sensors **17**(7), 1595 (2017). DOI 10.3390/s17071595
25. Pham, T.D., Park, C., Nguyen, D.T., Batchuluun, G., Park, K.R.: Deep learning-based fake-banknote detection for the visually impaired people using visible-light images captured by smartphone cameras. IEEE Access **8**, 63144–63161 (2020). DOI 10.1109/ACCESS.2020.2984019
26. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
27. Shafer, S.A.: Using color to separate reflection components. Color Research & Application **10**(4), 210–218 (1985)
28. Shreya, S.A., Siddique, M.A.I., Tasnim, S.: Real-time currency detection and voice feedback for visually impaired individuals (2025). ArXiv preprint
29. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G.: YOLOv10: Real-time end-to-end object

detection. In: Advances in Neural Information Processing Systems 37, pp. 107984–108011 (2024). DOI 10.52202/079017-3429
30. Wang, C.Y., Yeh, I.H., Liao, H.Y.M.: YOLOv9: Learning what you want to learn using programmable gradient information. In: Computer Vision – ECCV 2024, pp. 1–21 (2024). DOI 10.1007/978-3-031-72751-1_1
31. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11534–11542 (2020)
32. Wang, S., Toreini, E., Hao, F.: Anti-counterfeiting for polymer banknotes based on polymer substrate fingerprinting. IEEE Transactions on Information Forensics and Security **16**, 2823–2835 (2021). DOI 10.1109/TIFS.2021.3067440
33. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
34. Yang, L., Zhang, R.Y., Li, L., Xie, X.: SimAM: A simple, parameter-free attention module for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 11863–11874 (2021)
35. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU loss: Faster and better learning for bounding box regression. In: Proc. AAAI Conf. Artificial Intelligence, vol. 34, pp. 12993–13000 (2020)
36. Zuiderveld, K.: Contrast limited adaptive histogram equalization. In: Graphics Gems IV, pp. 474–485. Academic Press (1994)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf