

MAGE: Color-Invariant and Spatial Knowledge Distillation for Gastric Neoplasm Classification

Jiho Jun^{1,2}, Jeongwon Woo³, Jaemin Song^{1,2}, Thanh Bong Nguyen²,
Dong-heon Yeon^{1,2}, Donghoon Kang⁴, Jae-Myung Park⁴, Sung-Jea Ko^{1,2}, and
Kwang-Hyun Uhm⁵ *

¹ Korea University, Korea

² MEDAI, Korea

³ Seoul National University, Korea

⁴ The Catholic University of Korea, Seoul ST. Mary's Hospital Korea

⁵ Gachon University, Korea
khuhm@gachon.ac.kr

Abstract. Accurate differentiation between gastric adenoma and carcinoma during endoscopy is critical for clinical decision-making. Yet, this task is highly challenging due to high inter-class similarity and ambiguous boundaries between the two classes. Existing ROI-based classification methods often suffer from detection/segmentation error propagation and loss of surrounding global context. In contrast, full-image classification lacks the necessary spatial focus. Furthermore, we observe that deep neural networks gravitate towards domain-specific texture biases (e.g. bleeding, lighting artifacts), often causing models to predict based on spurious correlations instead of intrinsic morphological features. To address these limitations, we propose a novel framework, Masked Achromatic Guidance Expert (MAGE). During training, we introduce an auxiliary local expert branch trained on masked achromatic views of the neoplasm. By suppressing background context and color, this branch is forced to learn highly discriminative, purely structural features. We then employ a dual-objective distillation strategy, transferring both classification logits and spatial attention maps to provide implicit spatial supervision to the main branch that receives full WLI as input. This dual-objective distillation forces the model to ground its predictions in morphology rather than relying on shortcuts, while still retaining clinically relevant color cues. At inference time, our deployable model operates on images without annotated masks, ensuring real-time deployability. Extensive experiments on a clinical gastric endoscopy dataset show that our method significantly outperforms existing detection-based methodologies (e.g. YOLO) and classification-based methodologies (e.g. Swin-Transformer), providing not only superior classification performance but also interpretable attention maps for clinical reliability.

Keywords: Neoplasm classification · Gastric Endoscopy · Distillation · Adenoma · Adenocarcinoma · Explainability

* Corresponding author.

1 Introduction

Distinguishing gastric adenoma from carcinoma in upper gastrointestinal endoscopy is clinically important because it directly affects treatment strategy and prognosis in early gastric neoplasia [2]. However, this differential diagnosis remains a challenge even for experts. In a consecutive cohort of superficial elevated gastric neoplasms ($n=93$), diagnosis using C-WLI achieved 74% diagnostic accuracy, and implementing M-NBI improved accuracy to 92% [10]. However, this kind of enhanced imaging requires additional resources and could even become substantially harder in clinically ambiguous cases. Furthermore, diagnostic accuracy drops to 55.9% (C-WLI) / 58.1% (C-WLI + M-NBI) in deliberately matched difficult cases (50 vs. 50; 14 endoscopists) [16]. These findings highlight the intrinsic difficulty of fine-grained adenoma-carcinoma discrimination, where neoplasms exhibit highly similar appearances and ambiguous boundaries.

Recent deep learning methods for GI endoscopy largely target coarse discrimination (e.g., carcinoma vs. background), while fine-grained benign/precancerous vs. carcinoma classification remains under-addressed [8]. Detection based models show satisfactory performance in localizing neoplasms from background, but have limitations in diagnosis, especially for hard cases [17]. Conversely, classification models suffer from the black box problem; in our dataset, classification models often predict based on background regions, rather than the ROI itself (see Fig. 2). Furthermore, ROI-based pipelines deploying classification models on proposed ROI suffer from localization error propagation, while full-image classification preserves context but can dilute micro-patterns under background and acquisition artifacts [18].

We hypothesize that the key features distinguishing adenoma from carcinoma are primarily morphological and textural. While color does contain some diagnostic signal, it can also act as a fragile confounder that hinders deep neural networks from focusing on the primary morphological indicators of malignancy. This hypothesis is consistent with magnification-based endoscopic diagnosis literature, where microsurface and microvascular patterns vary with histology and invasion depth [11, 7]. In parallel, recent GI imaging studies report that color manipulation/correction do not consistently improve CAD performance and can even increase false positives, suggesting that color information may be unstable or task-dependent [1].

Based on this motivation, we propose **MAGE** (Masked Achromatic Guidance Expert), a framework utilizing a training-time auxiliary expert that learns color-invariant structural cues from a ROI-focused achromatic view (inside-mask grayscale; outside-mask zero). This achromatic expert demonstrated the highest diagnosis performance utilizing ground truth annotated masks, but cannot be deployed at inference time due to the lack of masks. Therefore, we transfer ROI-centric structural knowledge and spatial attention to a deployable model that receives full images as input via soft-target distillation and spatial attention transfer [3, 19].

In summary, our contributions are threefold: (1) a training-only auxiliary expert that promotes color-invariant morphological learning via a masked achro-

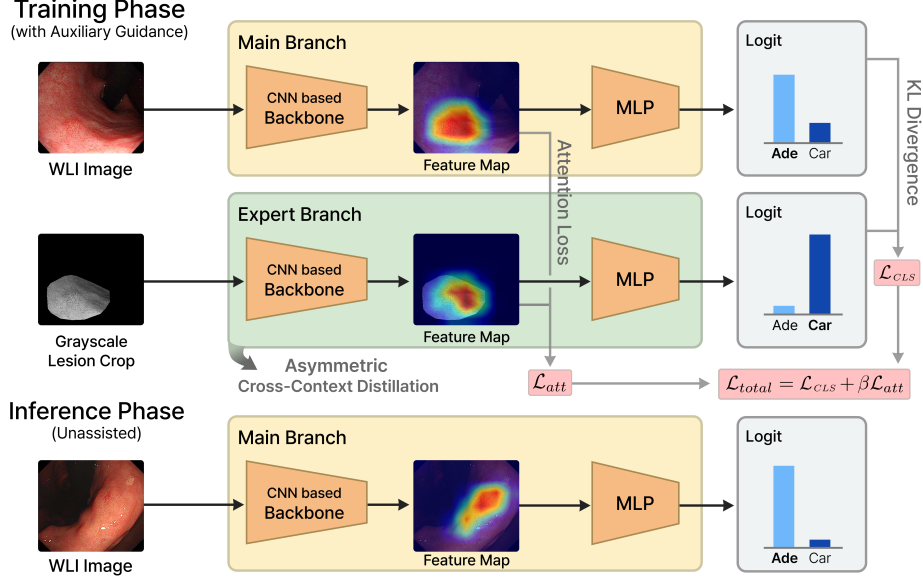


Fig. 1: Overview of the proposed dual-branch training framework.

matic input; (2) a spatial distillation framework that transfers ROI-centric structural cues to a deployable full-image classifier [3, 19]; and (3) consistent improvements over strong detection-and classification-based baselines and domain-pretrained backbones, supported by robustness and explainability analyses.

2 Proposed Method

During the training phase, raw WLI $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, annotated masks $\mathbf{m} \in \{0, 1\}^{H \times W}$, and binary labels $y \in \{0, 1\}$ are available. During the inference phase, only the images are available. We define the model’s components as the vision encoder E , feature map $f = E(\mathbf{x})$, classification head C , and output logit $z = C(f)$.

2.1 Masked Achromatic Expert

The Masked Achromatic Expert serves as an auxiliary branch utilizing privileged information to learn color invariant features in the neoplasm area. Because the masked region inherently contains the highest diagnostic signal-to-noise ratio within the image, this constraint forces the expert’s feature map to activate within the ROI. Concurrently, training on achromatic views forces the expert to learn robust, color invariant features in a domain where color is often a confounding factor. This training scheme transfers orthogonal training signals to

the deployable main model which receives full WLI as input. Given a segmentation mask $\mathbf{m}$, we construct a masked-achromatic input $\tilde{\mathbf{x}} = \mathbf{m} \odot \text{Gray}(\mathbf{x})$, where $\text{Gray}(\cdot)$ is a luma conversion and the grayscale channel is replicated to 3 channels for the vision encoder. We denote the expert’s output logits as $z_E = C_E(E_E(\tilde{\mathbf{x}}))$.

2.2 Main model

We transfer knowledge from the expert to the main model via knowledge distillation and spatial attention transfer. During training, the expert is kept frozen and produces a feature map and logit from the masked-achromatic input: $f_E = E_E(\tilde{\mathbf{x}})$ and $z_E = C_E(f_E)$. In parallel, the main model processes the raw WLI to obtain $f_M = E_M(\mathbf{x})$ and $z_M = C_M(f_M)$. Based on these outputs, we optimize the main model with the composite objective $\mathcal{L} = \mathcal{L}_{CLS} + \beta \mathcal{L}_{att}$.

Classification Loss . $\mathcal{L}_{CLS}$ consists of a weighted average between the BCE loss against the ground truth class label ($\mathcal{L}_{BCE}(z_M, y)$) and the KL divergence between the expert’s output logits on the image and the model’s output logits ($\tau^2 \mathcal{D}_{KL}(p_E^{(\tau)} \| p_M^{(\tau)})$), where $p_M^{(\tau)} = \sigma(\frac{z_M}{\tau})$, and $p_E^{(\tau)} = \sigma(\frac{z_E}{\tau})$, weighted by α and using temperature τ .

Attention Loss , $\mathcal{L}_{att}$. In order to transfer spatial attention to the main model, we collapse the channel dimension and normalize the feature map using $A(f) : \mathbb{R}^{B \times C \times H \times W} \rightarrow \mathbb{R}^{B \times H \times W}$. Setting pixels outside the mask to zero restricts expert activations within the ROI. The main model is regularized by squared Frobenius distance . Using the expert’s attention map as a supervision signal implicitly leads the main model to effectively localize neoplasms from the full image. This term is weighted by β , which controls the strength of the attention regularization.

Together, these terms transfer the expert’s privileged diagnostic knowledge while encouraging neoplasm localization from the full image. Accordingly, the total loss is defined as:

$$\mathcal{L} = \alpha \mathcal{L}_{BCE}(z_M, y) + (1 - \alpha) \tau^2 \mathcal{D}_{KL}(p_E^{(\tau)} \| p_M^{(\tau)}) + \beta \|A(f_M) - A(f_E)\|_F^2 \quad (1)$$

2.3 Rationale for Convolutional Architecture

While Vision Transformers (ViTs) are widely adopted in endoscopy tasks, we intentionally opted for a Convolutional Neural Network (CNN) backbone. This decision is motivated by dataset constraints and the inherent characteristics of medical imaging. By design, ViTs lack strong inductive biases—such as translation invariance and locality—and therefore require large-scale datasets for effective generalization. Given our dataset, CNN architectures offer crucial regularization via local inductive biases, facilitating stable training without the need for extensive pre-training. Furthermore, our proposed distillation framework relies heavily on precise spatial alignment between the intermediate feature maps of

the main and expert branches [19]. The inherent hierarchical, grid-like feature representations of CNNs allow for a more direct and efficient computation of spatial attention loss. Therefore, considering both the visual characteristics of WLI endoscopy and the context-aware knowledge distillation mechanism, our framework is fundamentally tailored for CNN-based architectures.

3 Experiments and Results

3.1 Experimental Setup

Task. Binary classification of gastric neoplasia (adenoma vs. carcinoma) from WLI. **Datasets.** We utilized a private WLI dataset collected from Seoul St. Mary’s Hospital, Korea (IRB: KC20RISI0503), with annotation masks for in-domain training/evaluation, partitioned into patient-disjoint training (840 patients; 2443 images; 1,222 adenoma / 1,221 carcinoma), validation (180 patients; 488 images; 240/248), and test (181 patients; 512 images; 252/260) splits. For external testing, we evaluate our methodology on the PICCOLO [14] dataset (40 patients; 1,727 images; 1,454/ 273), excluding non-neoplastic cases (e.g., hyperplasia). **Metrics.** AUC, F1, Precision, and Recall with $\sigma(z_S)$ and a 0.5 threshold unless otherwise stated. Models were trained on an RTX3090 (Adam/AdamW, cosine annealing $1e-4$ to $1e-6$, batch size 32, 224×224 inputs) utilizing $\tau = 4.0$ and $\beta = 300$. All benchmarks used random-flip augmentation only.

3.2 Benchmarks

Table 1: Main Results (20-run average). [†] Classification-head metrics.

Method	AUC	F1	Prec.	Rec.
GastroNet-5M [5] Res50 pretrained	0.744 ± 0.022	0.704 ± 0.012	0.683 ± 0.041	0.734 ± 0.055
PolyDSS [13] [†]	0.750 ± 0.013	0.700 ± 0.013	0.698 ± 0.015	0.704 ± 0.028
YOLO26-L [4] [†]	0.737 ± 0.014	0.691 ± 0.018	0.693 ± 0.016	0.692 ± 0.017
UML [12] [†]	0.639 ± 0.038	0.672 ± 0.012	0.548 ± 0.028	0.879 ± 0.080
SAM (ViT-B) [6]+ Swin-Tiny[9]	0.719 ± 0.020	0.700 ± 0.014	0.586 ± 0.035	0.879 ± 0.061
U-Net (Res50)+ Swin-Tiny [9]	0.731 ± 0.012	0.718 ± 0.011	0.615 ± 0.036	0.870 ± 0.060
Ours (MAGE, $\alpha = 0.25$)	0.786 ± 0.010	0.733 ± 0.010	0.673 ± 0.017	0.806 ± 0.035

Table 1 compares our approach with representative baselines and recent Gastronet 5M pretrained backbones. To ensure fair comparison across heterogeneous architectures, we report image-level classification metrics: joint models

use their classification head outputs, and two-stage pipelines(e.g. SAM + Swin-Transformer) classify ROIs localized by segmentation. We report the average and standard deviation across 20 runs. In particular, MAGE improves F1 and AUC over the strongest classification baselines, indicating more reliable overall decision quality under the same WLI-only setting.

3.3 Explainability and Robustness

Explainability: Grad-CAM / Attention Concentration We analyze model explainability using Grad-CAM [15] and quantify how well the resulting attribution maps align with neoplasm masks (Table 2). Specifically, we report: (i) *Attention-in-Mask (AiM)*, the fraction of total attribution mass that falls inside the ground-truth mask; (ii) the *Pointing Game Accuracy(PGA)* [20], which checks whether the maximum-saliency point lies within the neoplasm region; and (iii) a *saliency Dice* score computed between binarized saliency maps and the ground-truth masks, where the binarization threshold is selected on the validation set to maximize the score. Overall, MAGE yields more localized and mask-consistent explanations, indicating that the model relies more on neoplasm morphology/structure rather than spurious global color patterns. Notably, its explanation quality is comparable to that of a strong baseline trained directly on segmentation objectives. Fig. 2 presents qualitative Grad-CAM examples.

Table 2: Grad-CAM + Attention Evaluation. [†] YOLO reports standard segmentation Dice (not saliency mDice) and is shown for reference.

Method	AiM	PGA	Saliency mDice
Baseline (EffNet)	0.3176	0.4355	0.3225 @ $t = 0.15$
GastroNet pretrained ResNet50	0.3727	0.5117	0.3600 @ $t = 0.10$
Swin-Transformer Tiny	0.2023	0.3398	0.2492 @ $t = 0.10$
Ours ($\alpha = 0.5$)	0.4876	0.6758	0.4751 @ $t = 0.20$
Ours ($\alpha = 0.25$)	0.5303	0.7324	0.5308 @ $t = 0.20$
Ours ($\alpha = 0$)	0.5484	0.7734	0.5489 @ $t = 0.20$
YOLO26-Seg Large [†] (<i>for reference</i>)	0.6313	0.7695	0.6361 (Dice)

Color Robustness: Hue Shift / Color Perturbation Gastric endoscopy images often exhibit non-diagnostic appearance variations that can shift the observed color tone and illumination. To test the model’s robustness against color variations, we evaluate performance on images subject to synthetic hue rotations. Specifically, we apply a range of hue shifts to the validation set and compare our model against a baseline (EffNet-v2-s) and benchmark model(ResNet50 pretrained on GastroNet-5M, finetuned on our dataset). As shown in Fig. 3, our method shows superior overall performance compared to both the EfficientNet trained baseline and GastroNet-5M pretrained ResNet50 baseline. Especially,

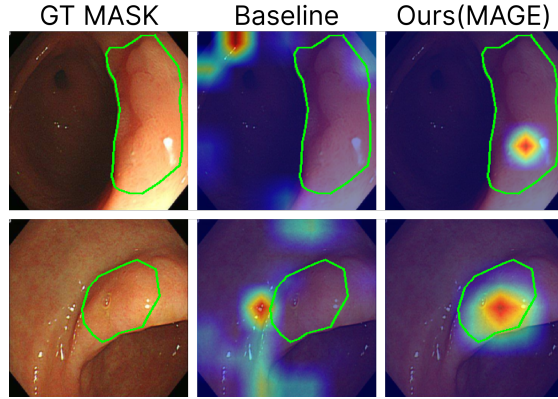


Fig. 2: Examples of Grad-CAM, GT, Baseline(EfficientNet), MAGE.

while the naively trained EfficientNet degrades significantly over positive hue shifts, our MAGE model ($\alpha = 0$) shows robustness on par with the GastroNet pretrained baseline. Although ($\alpha = 0$) maximizes color robustness and spatial alignment, we utilize ($\alpha = 0.25$) in benchmarks to optimally balance label-guided accuracy.

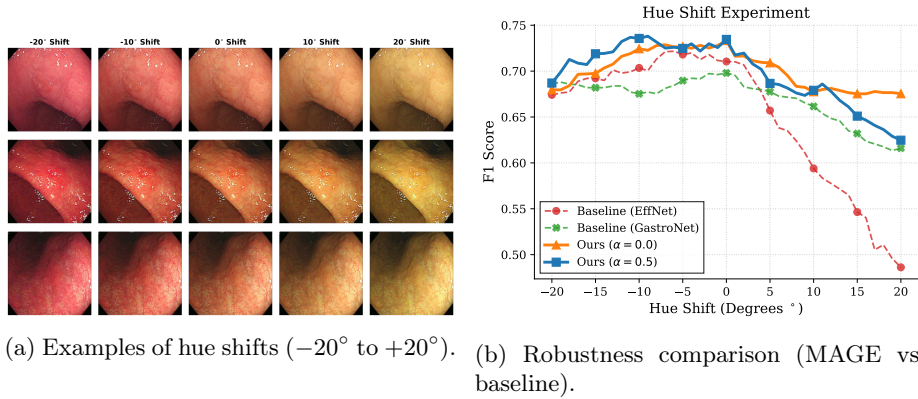


Fig. 3: Color robustness under synthetic hue perturbations.

3.4 External Dataset: PICCOLO (Cross-domain Generalization)

We further evaluate cross-domain generalization on PICCOLO [14], an external colon endoscopy dataset. We test whether our training methodology—focused on color-invariant, neoplasm-centric morphological features—is transferable beyond the gastric domain. Due to the limited number of unique neoplasms in

Table 3: PICCOLO results.

Method	AUC	mF1	mPrecision	mRecall
GastroNet-5M pretrained [5] ResNet50	0.9110	0.8053	0.8269	0.6717
Swin-Tiny [9] (ImageNet pretrained)	0.9293	0.8534	0.9004	0.8440
U-Net (Res50) + Swin-Tiny [9]	0.9096	0.7858	0.7975	0.8099
YOLO 26 Large [4]	0.7816	0.7521	0.8566	0.6959
Ours (MAGE)	0.9434	0.8173	0.9298	0.7863

PICCOLO, we report metrics averaged over a 10-repetition 3-fold cross validation. We report AUC and macro-averaged F1/Precision/Recall to account for class imbalance. As shown in Table 3, our method achieves the highest AUC, supporting the overall robustness under the change of the domain.

3.5 Ablation Study

The highest performance was achieved by pairing a cropped grayscale Expert, which focuses on intrinsic neoplasm structure, with an RGB Main branch that leverages morphological and color cues. The Main branch exhibits a limited ability to distinguish neoplasms in grayscale images. However, for the expert receiving masked inputs, training on grayscale inputs exhibited a higher classification performance compared to RGB inputs. These findings suggest that an *asymmetric design* intentionally constrains spatial and color information to force the model to learn structural, color-invariant features. $\alpha = 0.25$ was used in training the main inputs.

Table 4: Ablation Study on Knowledge Distillation Configurations.

Expert Configuration				Main Input : Grayscale		Main Input : RGB	
RGB	Gray	Full	Crop	AUC	F1	AUC	F1
✓		✓		0.7483	0.7081	0.7738	0.7192
✓			✓	0.7082	0.6872	0.7748	0.7177
	✓	✓		0.7524	0.7147	0.7767	0.7235
	✓		✓	0.7115	0.6940	0.7860	0.7330

4 Conclusion

Contribution. We presented MAGE, a dual-branch training framework that promotes color-invariant morphological representations for gastric neoplasm diagnosis. Our key contribution is a training-only masked grayscale expert that suppresses color shortcuts, and distills its structural knowledge into a deployable full-image model. As a result, MAGE improves classification performance and robustness under color/domain variability.

Limitations and Future Work. Our study focuses on classification and robustness without explicitly training a dedicated detection/segmentation module, which may limit direct localization outputs required in some clinical workflows. A natural next step is to leverage the learned classifier for lightweight localization (e.g., CAM-based or weakly supervised cues) without incurring the cost of a full segmentation system. In addition, broader external validation, particularly multi-center and device-diverse cohorts, remains essential to establish generalization and clinical utility. Code and weights will be publicly released.

Acknowledgments. This work was supported by the Technology development Program(RS-2024-00460263) funded by the Ministry of SMEs and Startups(MSS, Korea). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2026-25472209). This research was supported by the “Advanced GPU Utilization Support Program” funded by the Government of the Republic of Korea (Ministry of Science and ICT).

Disclosure of Interests. S.-J. Ko is the CEO of MEDAI. J. Jun, J. Woo, J. Song, T. B. Nguyen, and D.-h. Yeon are employees of MEDAI. The remaining authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agossou, B.E., Pedersen, M., Raja, K., Vats, A., Floor, P.A.: Influence of color correction on pathology detection in capsule endoscopy. arXiv preprint arXiv:2502.00076 (2025)
2. Fujishiro, M.: Advanced diagnostic and therapeutic endoscopy for early gastric cancer. *Cancers* **16**(5), 1039 (2024). <https://doi.org/10.3390/cancers16051039>, <https://pubmed.ncbi.nlm.nih.gov/38473395/>
3. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *CoRR* **abs/1503.02531** (2015), <https://arxiv.org/abs/1503.02531>
4. Jocher, G., Qiu, J.: Ultralytics yolo26. <https://github.com/ultralytics/ultralytics> (2026), license: AGPL-3.0
5. Jong, M.R., et al.: Gastronet-5m: A multicenter dataset for developing foundation models in gastrointestinal endoscopy. *Gastroenterology* (2026). <https://doi.org/10.1053/j.gastro.2025.07.030>, <https://pubmed.ncbi.nlm.nih.gov/40749857/>
6. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). <https://doi.org/10.1109/ICCV51070.2023.00371>
7. Kurumi, H., Nonaka, K., Ikebuchi, Y., Yoshida, A., Kawaguchi, K., Yashima, K., Isomoto, H.: Fundamentals, diagnostic capabilities and perspective of narrow band imaging for early gastric cancer. *Journal of Clinical Medicine* **10**(13), 2918 (2021). <https://doi.org/10.3390/jcm10132918>, <https://pubmed.ncbi.nlm.nih.gov/34209939/>
8. Liu, J., Li, D., Zhuo, Y., Zhang, S.: Deep learning for detecting early gastric cancer with white-light endoscopy: a systematic review and meta-analysis. *Frontiers in Artificial Intelligence* **9**, 1734591 (2026). <https://doi.org/10.3389/frai.2026.1734591>

9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021). <https://doi.org/10.1109/ICCV48922.2021.00986>
10. Maki, S., Yao, K., Nagahama, T., et al.: Magnifying endoscopy with narrow-band imaging is useful in the differential diagnosis between low-grade adenoma and early cancer of superficial elevated gastric lesions. *Gastric Cancer* **16**, 140–146 (2013). <https://doi.org/10.1007/s10120-012-0160-7>
11. Ok, K.S., Kim, G.H., Park, D.Y., Lee, H.J., Jeon, H.K., Baek, D.H., Lee, B.E., Song, G.A.: Magnifying endoscopy with narrow band imaging of early gastric cancer: Correlation with histopathology and mucin phenotype. *Gut and Liver* **10**(4), 532–541 (2016). <https://doi.org/10.5009/gnl15364>, <https://pubmed.ncbi.nlm.nih.gov/27021504/>
12. Ren, K., Zou, K., Liu, X., Chen, Y., Yuan, X., Shen, X., Wang, M., Fu, H.: Uncertainty-informed mutual learning for joint medical image classification and segmentation. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Lecture Notes in Computer Science, vol. 14223, pp. 35–45. Springer Nature Switzerland, Vancouver, BC, Canada (2023). https://doi.org/10.1007/978-3-031-43901-8_4
13. Saad, A.I., Maghraby, F.A., Badawy, O.M.: Polydss: computer-aided decision support system for multiclass polyp segmentation and classification using deep learning. *Neural Computing and Applications* **36**, 5031–5057 (2024). <https://doi.org/10.1007/s00521-023-09358-3>, <https://link.springer.com/article/10.1007/s00521-023-09358-3>
14. Sánchez-Peralta, L.F., Pagador, J.B., Picón, A., et al.: Piccolo white-light and narrow-band imaging colonoscopic dataset: A performance comparative of models and datasets. *Applied Sciences* **10**(23), 8501 (2020). <https://doi.org/10.3390/app10238501>, <https://www.mdpi.com/2076-3417/10/23/8501>
15. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). <https://doi.org/10.1109/ICCV.2017.74>
16. Tamura, N., Sakaguchi, Y., Furutani, W., et al.: Magnifying endoscopy with narrow-band imaging is useful in differentiating gastric cancer from matched adenoma in white light imaging. *Scientific Reports* **12**, 8349 (2022). <https://doi.org/10.1038/s41598-022-12315-0>, <https://www.nature.com/articles/s41598-022-12315-0>
17. Tudela, Y., Majó, M., de la Fuente, N., et al.: A complete benchmark for polyp detection, segmentation and classification in colonoscopy images. *Frontiers in Oncology* (2024). <https://doi.org/10.3389/fonc.2024.1417862>
18. Yin, T.K., et al.: Endoscopy artefact detection by deep transfer learning of baseline models. *Sensors* (2022), <https://pmc.ncbi.nlm.nih.gov/articles/PMC9582060/>
19. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: *International Conference on Learning Representations (ICLR)* (2017), <https://arxiv.org/abs/1612.03928>
20. Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. *arXiv preprint arXiv:1608.00507* (2016), <https://arxiv.org/abs/1608.00507>