Computer Methods and Programs in Biomedicine 277 (2026) 109260

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Enhancing accuracy and explainability in colorectal lesion classification with attention-supervised Vision Transformers

Luca Carlini [a,*], Luca Di Stefano [a], Chiara Lena [a], Davide Massimi [b], Tommy Rizkala [b], Cesare Hassan [b], Elena De Momi [a]

[a] Dipartimento di Elettronica, Informazione, Bioingegneria (DEIB), Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milano, 20133, Italy
[b] IRCCS Humanitas Research Hospital, Via Manzoni, 56, Rozzano (MI), 20089, Italy

ARTICLE INFO

Code link: https://github.com/LucaCarlini/SUNDatasetPretraining

Keywords:
Vision transformers
Attention supervision
Paris classification
Trustworthy AI
Colorectal lesion classification

ABSTRACT

Objective: Accurate assessment of colorectal lesion morphology during colonoscopy is essential for guiding treatment and estimating cancer risk. The Paris classification is widely adopted for this purpose but suffers from substantial inter-observer variability, while Vision Transformers (ViTs) can base their decisions on diffuse, off-lesion attention patterns that are hard to interpret. This study investigates whether directly supervising ViT attention maps with expert lesion annotations can concurrently improve Paris classification performance and model explainability.

Method: We propose a Lesion-Focused Attention Loss ($\mathcal{L}_{\mathrm{LFA}}$), an attention-supervised pretraining objective that uses expert polyp bounding boxes to focus last-layer [CLS] attention on annotated lesion regions, followed by standard cross-entropy fine-tuning. $\mathcal{L}_{\mathrm{LFA}}$ is applied to six ViT architectures and evaluated on the public SUN dataset for binary (0-I vs. 0-II) and three-class (0-Ip, 0-Is, 0-IIa) Paris classification. Performance is assessed using frame-wise accuracy and the AttIn, we additionally perform an ablation study against a Grad-CAM consistency baseline.

Results: Attention-supervised pretraining yields consistent gains in both accuracy and lesion-focused attention. Across the six ViTs, adding $\mathcal{L}_{\mathrm{LFA}}$ improves three-class accuracy by up to 7 percentage points. In a detailed ablation on ViT-B/16, $\mathcal{L}_{\mathrm{LFA}}$ outperforms a Grad-CAM consistency baseline by about 5–13 percentage points across the 2-class and 3-class tasks, and χ^2 tests confirm a significant association between high AttIn and correct predictions.

Conclusion: Direct supervision of ViT attention with $\mathcal{L}_{\mathrm{LFA}}$ leverages expert knowledge to jointly boost Paris classification accuracy and spatial interpretability, and compares favourably with Grad-CAM–based explanation regularisation. The source code and dataset splits are publicly available at https://github.com/LucaCarlini/SUNDatasetPretraining.

1. Introduction

Colorectal cancer is a leading cause of cancer-related morbidity and mortality, and colonoscopy remains the gold standard for its prevention, enabling direct inspection of the colonic mucosa and resection of precancerous lesions in a single session [1,2]. Artificial intelligence (AI) systems now support endoscopists with real-time computer-aided detection (CADe) and diagnosis (CADx) of colorectal lesions [3,4], but are often deployed as black boxes with limited transparency. Morphology is typically assessed using the Paris classification [5], which stratifies superficial lesions into polypoid (0-Ip, 0-Is) and non-polypoid (0-IIa, 0-IIb, 0-IIc, 0-III) categories linked to submucosal invasion risk and treatment strategy. However, Paris classification shows substantial inter-observer variability [6,7], underscoring the need for AI systems that are both accurate and interpretable.

Deep learning is now the standard for automated colorectal lesion classification, with Convolutional Neural Networks (CNNs) achieving strong CADe/CADx performance. Vision Transformers (ViTs) have recently emerged as competitive or superior backbones [8,9], leveraging self-attention to capture long-range context and exposing attention weights that can be visualised as intrinsic explanations [10]. Krenzer et al. [8] reported 89.35% frame-wise Paris accuracy on SUN [11,12] using a ViT-L/16, outperforming strong CNN baselines, but relying on private training data and without quantitative analysis of attention behaviour. More broadly, raw ViT attention is highly sensitive to the pretraining objective and can be diffuse or clinically implausible [13,

* Corresponding author.
E-mail address: luca.carlini@polimi.it (L. Carlini).

https://doi.org/10.1016/j.cmpb.2026.109260
Received 22 September 2025; Received in revised form 11 January 2026; Accepted 19 January 2026
Available online 21 January 2026

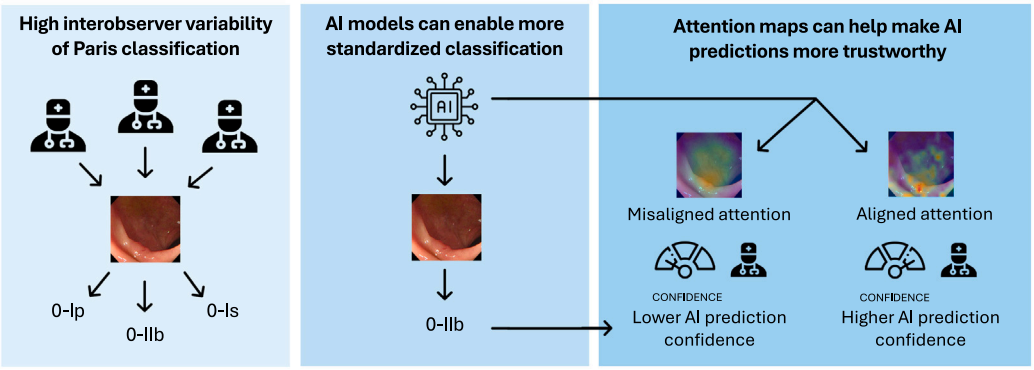


Fig. 1. Motivation for attention-supervised ViTs. High inter-observer variability in Paris morphology assessment yields inconsistent labels for the same lesion (left). AI models can provide a more standardised Paris classification (middle). Attention maps indicate whether the model attends to the clinically relevant lesion region, where misaligned attention is associated with lower surgeon confidence in the model output and aligned attention supports higher surgeon confidence and more trustworthy predictions (right).

14], while the XAI literature shows that post-hoc saliency methods such as Grad-CAM may not faithfully capture model reasoning [10,15–20] and are rarely evaluated in realistic clinical workflows [21–23]. Fig. 1 summarises these challenges, illustrating both the inter-observer variability in Paris morphology labels and the dispersed, off-lesion focus of baseline ViT attention maps. Together, these dual sources of uncertainty limit the reproducibility of Paris classification and constrain the interpretability of AI decisions, since predictions may rely on image regions that are not clinically meaningful, whereas attention maps provide a mechanism to assess whether a given output is likely reliable by verifying that model focus aligns with the lesion.

Motivated by the empirical observation quantified in Section 4 that baseline ViT frames with naturally lesion-focused attention achieve higher accuracy than those with diffuse or off-lesion maps, we explicitly inject expert spatial knowledge on lesion location into ViT attention. We introduce a Lesion-Focused Attention Loss $\mathcal{L}_{\mathrm{LFA}}$, an attention-supervised pretraining objective that uses expert polyp bounding boxes to steer the last-layer `[CLS]` attention towards the annotated lesion region before standard Paris classification fine-tuning. This design directly supervises intrinsic attention maps with expert knowledge to concurrently boost performance and interpretability: it embeds spatial clinical priors into the backbone and turns lesion-focused attention into a meaningful indicator of prediction reliability. Across multiple ViT backbones, attention-supervised pretraining increases the fraction of attention inside the lesion, improves Paris classification accuracy, and strengthens the association between high attention–lesion overlap and correct predictions.

2. Related work

2.1. Lesion classification

Colorectal lesion classification has mainly been addressed with CNN-based CADe/CADx systems, covering tasks from binary adenoma detection to multi-class histological or morphological grading, often with high reported accuracies but predominantly on private datasets [24–29]. Only a few studies directly target Paris morphology. Bour et al. grouped Paris types into three risk strata using ResNet-50 and other CNNs, achieving 87.1% accuracy on 785 images [30]. Krenzer et al. combined polyp detection with a ViT-L/16 classifier and reached 89.35% frame-wise Paris accuracy on SUN [11,12], outperforming strong CNN baselines [8]. However, their training data are private and attention behaviour is not quantified. Overall, prior work rarely analyses whether model focus actually coincides with the annotated lesion. In contrast, we use SUN as a fully public benchmark and explicitly measure the spatial alignment between ViT `[CLS]` attention and expert-delineated polyp bounding boxes.

2.2. Attention, region supervision and interpretability

ViTs expose attention weights that are often interpreted as intrinsic explanations [10,13,14], but several studies show that attention structure is highly sensitive to the pretraining objective and can be diffuse or clinically implausible even when accuracy is good [13,14]. Parallel work has explored making explanations trainable. For CNNs, Ismail et al. sharpened saliency maps by penalising changes under input masking [31]. For ViTs, LGMViT supervises attention with foreground masks and an auxiliary localisation head, improving both alignment and accuracy on brain and liver imaging [32]. These methods show that weak spatial supervision can shape model focus, but they either require dense labels or do not treat attention–lesion overlap as a quantitative confidence signal.

Yu et al. recently proposed a Location-guided Lesions Representation Learning (LLRL) framework based on image generation for plant leaf disease severity assessment [33]. Their approach combines an image generation network (IG-Net), which uses a diffusion model to synthesise diseased leaves from healthy ones, with a location-guided lesion representation learning network (LGR-Net) that contrasts healthy/diseased pairs to learn lesion-focused features, and a hierarchical fusion network for final severity prediction. While conceptually related to our goal of embedding lesion priors, LLRL relies on training a high-capacity generative model on a wide variety of lesion appearances and backgrounds. This presupposes substantial lesion diversity, which is not available in public colorectal endoscopy datasets making, a direct application of LLRL impractical in our setting. This presupposes substantial lesion diversity and the availability of healthy–diseased image pairs from the same scene, conditions that are not met by public colorectal endoscopy datasets such as SUN (which contain many frames but only 100 distinct lesions and no paired healthy/diseased views). For these reasons, a direct implementation of LLRL on SUN would require substantial redesign and would not constitute a fair or informative baseline, so we do not include it in our experimental comparison.

In natural-image classification, explanation regularisation is often based on post-hoc Grad-CAM. Lee and Cho [34] introduced a Grad-CAM consistency loss in a semi-supervised setting, where a classifier is trained not only to produce consistent label predictions for an image and its augmented counterpart, but also to yield similar Grad-CAM maps at a chosen target layer. Concretely, they generate a pseudo-label from the original image, apply stochastic augmentations (e.g. cropping, rotation, colour jitter) to obtain a perturbed view, and then minimise a combined objective consisting of: a supervised cross-entropy term on labelled data, a pseudo-label consistency term between original and augmented predictions on unlabelled data, and an ℓ_2 loss between the Grad-CAM heatmaps computed on the two views. By attaching the Grad-CAM module to different ResNet blocks, they show

that constraining explanation consistency at deeper layers can improve generalisation on CIFAR-10. Zhu et al. extended this idea to person re-identification [35]. Other approaches operate on geometric regions [36] or distil attention from a teacher [37]. These methods generally ignore ground-truth object masks and are designed for CNNs or generative models rather than ViT classifiers. In this work we explicitly re-implement and compare against the Grad-CAM consistency method of Lee and Cho as a strong, explanation-regularised baseline.

2.3. Positioning of the proposed method

The proposed approach lies at the intersection of large-scale ViT pretraining, explanation-regularised learning and clinically grounded region supervision. Recent work such as DINOv2 shows that strong pretraining substantially improves ViT performance and robustness [38,39], motivating us to treat attention supervision as an additional pretraining signal rather than as a post-hoc constraint. Grad-CAM–based consistency methods further indicate that constraining explanation maps during training can influence predictive behaviour [34,35], but they operate on post-hoc saliency rather than intrinsic attention. In colonoscopy, existing ViT-based systems typically use attention maps only for qualitative visualisation and rarely quantify their alignment with annotated polyp regions [8,10], leaving the link between attention patterns and diagnostic reliability largely unexplored.

Compared with the LLRL framework of Yu et al. [33], which embeds lesion priors through a diffusion-based image generation network and a location-guided representation learner trained on synthetic healthy/diseased pairs, our method is architecturally simpler and better suited to public colorectal datasets such as SUN, which contain many frames but only a small number of distinct lesions. Our Lesion-Focused Attention Loss acts directly on intrinsic `[CLS]` attention in ViT classifiers rather than on Grad-CAM maps [34,35] or distilled teacher attention [37], leverages expert polyp bounding boxes instead of generic regions or dense masks [32,36], and is used as a standalone pretraining loss without auxiliary localisation heads [32]. We then evaluate both Paris classification accuracy and the statistical relationship between attention–lesion overlap and prediction correctness, benchmarking our method directly against Grad-CAM consistency [34] on the same public colonoscopy dataset.

3. Materials and methods

3.1. SUN dataset

The SUN dataset [11,12] is a publicly available collection of colonoscopic video frames for computer-aided detection and classification in gastrointestinal endoscopy. It comprises data from 99 patients (71 male, 28 female, median age 69 years, interquartile range 58–74) and 100 colorectal lesions, each polyp annotated with a bounding box and labelled according to the Paris morphological classification by expert endoscopists. All images were acquired using high-definition endoscopes (CF-HQ290ZI and CF-H290ECI, Olympus, Tokyo, Japan), and colonoscopies were recorded with a high-definition video recorder (IMH-10, Olympus). Every frame was initially annotated by three research assistants for polyp presence or absence, and polyp locations were delineated with bounding boxes. These annotations were then double-checked and, if necessary, corrected by two expert endoscopists with experience of more than 5000 colonoscopies. A frame was defined as positive for a polyp if more than half of the polyp appeared within the frame.

3.2. Dataset splitting

To obtain unbiased performance estimates, we split only the SUN dataset into case-level train/validation/test sets (81/9/10 cases), stratified to preserve the Paris morphology distribution across splits using the scikit-learn `train_test_split` function with stratification on the Paris label. We merged 0-Isp polyps into 0-Ip, reflecting their similar management (Table 1). We considered two classification scenarios: a 2-class task distinguishing polypoid (0-I) from non-polypoid (0-II) lesions, and a 3-class task with classes 0-Ip, 0-Is and 0-IIa, with lesion- and frame-level counts reported in Table 1. To mitigate frame-level class imbalance due to variable frames per lesion, we applied random undersampling of the over-represented polypoid class. In the 2-class setup we capped 0-I frames at 15,000/1,000/2,000 in train/validation/test, while retaining all 0-II frames. In the 3-class setup we analogously limited 0-Is frames to at most 1000 in the validation split, while keeping all 0-Ip and 0-IIa frames. This procedure yielded approximately balanced frame-level class proportions and only small differences in total frame counts between the 2-class and 3-class setups [40], and the exact image lists for each split, together with the splitting script, are provided in the public repository.

3.3. Attention-inside-lesion score (AttIn) definition

The intuition behind the attention-inside-lesion score (AttIn) is to obtain a quantitative measure of how much attention is concentrated on the lesion region. To this end, we compute the proportion of the `[CLS]` token's attention distribution that lies inside versus outside the lesion bounding box. This provides a direct way to assess whether the attention map is focused on the clinically relevant region when making a prediction.

To obtain a single-channel `[CLS]` attention map for each image, the last-layer `[CLS]` self-attention is averaged over all heads and reshaped into a 2D spatial grid. The resulting tensor $A_i \in \mathbb{R}^{H\times W}$ is bilinearly upsampled to a fixed resolution of 224×224 to match the scale of the bounding-box annotations. Since raw attention values differ in scale, each map is normalised in two steps: first by min–max scaling $\tilde{A}_i(x,y)$ (Eq. (1)) and then by $\mathcal{L}_1$ normalisation (Eq. (2)) so that the total mass sums to one, $\hat{A}_i(x,y)$:

$$\tilde{A}_i(x,y) = \frac{A_i(x,y) - \min_{u,v} A_i(u,v)}{\max_{u,v} A_i(u,v) - \min_{u,v} A_i(u,v) + \varepsilon}, \tag{1}$$

$$\hat{A}_i(x,y) = \frac{\tilde{A}_i(x,y)}{\sum_{u=1}^{H}\sum_{v=1}^{W} \tilde{A}_i(u,v)}, \qquad \sum_{x=1}^{H}\sum_{y=1}^{W} \hat{A}_i(x,y) = 1, \tag{2}$$

where $\varepsilon = 10^{-8}$ prevents numerical instability. As further shown in the ablation study, the $\mathcal{L}_1$ normalisation stabilises attention supervision and provides a more reliable guidance signal for the attention maps.

Given a bounding box $B_i = [x_1^{(i)}, y_1^{(i)}, x_2^{(i)}, y_2^{(i)}]$, we construct a binary mask M_i with value 1 inside the lesion region and 0 elsewhere:

$$M_i(x,y) = \begin{cases} 1, & x_1^{(i)} \le x \le x_2^{(i)} \text{ and } y_1^{(i)} \le y \le y_2^{(i)}, \\ 0, & \text{otherwise.} \end{cases}$$

The fractions of attention falling inside, representing the attention focus on the lesion, and outside the annotated region are then computed as in Eqs. (3) and (4):

$$\text{AttIn}_i = \sum_{x=1}^{H}\sum_{y=1}^{W} \hat{A}_i(x,y)\, M_i(x,y), \tag{3}$$

$$\text{AttOut}_i = \sum_{x=1}^{H}\sum_{y=1}^{W} \hat{A}_i(x,y)\,(1 - M_i(x,y)). \tag{4}$$

3.4. Attention-guided pretraining loss

The Lesion-Focused Attention Loss $\mathcal{L}_{\text{LFA}}$ for a single batch is defined as in Eq. (5):

$$\mathcal{L}_{\text{LFA}} = \frac{1}{N}\sum_{i=1}^{N} \text{AttOut}_i. \tag{5}$$

Minimising $\mathcal{L}_{\text{LFA}}$ encourages the model to focus its attention precisely on clinically relevant regions. Fig. 2 illustrates the overall training workflow.

Table 1
SUN dataset composition. Lesion- and frame-level distribution per Paris category in the train, validation and test splits. Lesion counts (top) are reported over 100 cases, frame counts (middle and bottom) correspond to the two-class (0-I vs. 0-II) and three-class (0-Ip, 0-Is, 0-IIa) setups after random undersampling, with percentages given relative to the row totals.

Level	Paris category	Train	Validation	Test
Lesion count (100 lesions)	0-Ip (0-I)	13 (16.0%)	2 (22.2%)	2 (20.0%)
	0-Is (0-I)	40 (49.4%)	4 (44.4%)	5 (50.0%)
	0-IIa (0-II)	28 (34.6%)	3 (33.3%)	3 (30.0%)
	Total	**81 (81.0%)**	**9 (9.0%)**	**10 (10.0%)**
Frame count 2-class (35,135 samples)	0-I	15000 (50.7%)	1000 (54.1%)	2000 (54.5%)
	0-II	14615 (49.3%)	849 (45.9%)	1671 (45.5%)
	Total	**29615 (84.3%)**	**1849 (5.3%)**	**3671 (10.5%)**
Frame count 3-class (30,426 samples)	0-Ip	6289 (27.3%)	978 (34.6%)	1579 (34.6%)
	0-Is	9440 (41.0%)	1000 (35.4%)	1312 (28.8%)
	0-IIa	7308 (31.7%)	849 (30.0%)	1671 (36.6%)
	Total	**23037 (75.7%)**	**2827 (9.3%)**	**4562 (15.0%)**

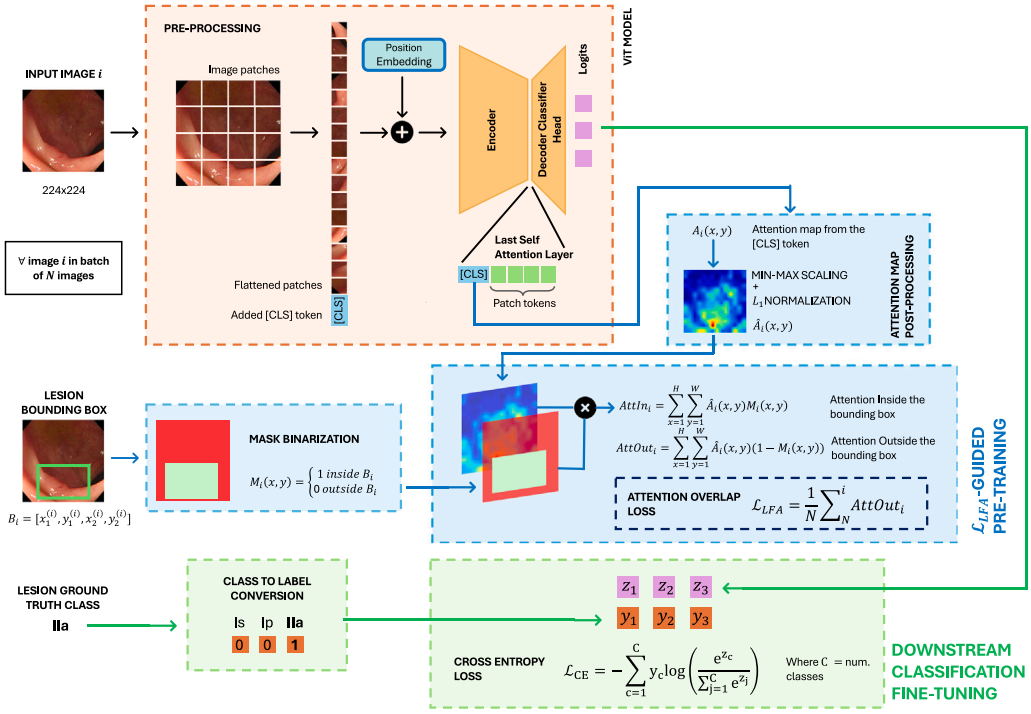


Fig. 2. Pretraining (blue) and classification training (green) pipelines. Each input frame is split into patches, prepended with a [CLS] token and passed through a ViT encoder. During $\mathcal{L}_{\text{LFA}}$-guided pretraining (blue arrows), the last-layer [CLS] attention map is min–max scaled and $\mathcal{L}_1$-normalised, combined with a binary mask from the lesion bounding box to compute AttIn/AttOut, and optimised with the Lesion-Focused Attention Loss $\mathcal{L}_{\text{LFA}}$. In the subsequent classification fine-tuning stage (green arrows), the pretrained model is trained using cross-entropy on Paris class labels.

3.5. Experimental setup

All experiments are conducted on the SUN dataset described in Sections 3.1–3.3, using the case-level train/validation/test splits and Paris classification tasks reported in Table 1. The evaluation metrics and statistical tests used throughout the study are detailed in Sections 3.7 and 3.8, respectively.

We evaluate the proposed attention-supervised strategy on six ViT backbones spanning different capacities, patch sizes and pretraining paradigms: ViT-B/16, ViT-L/16, ViT-B/32, ViT-L/32, DINOv2-B and DINOv2-L [38,39]. Here, *B* and *L* denote base and large model sizes, respectively, and the patch size (16 vs. 32) controls the spatial granularity of the token grid, with smaller patches preserving finer lesion detail at higher computational cost. DINOv2 backbones are included to assess whether self-distilled DINO-style pretraining generalises better than supervised ViT pretraining [38].

All experiments are run on a single NVIDIA A40 GPU (40 GB VRAM). The same hardware, dataset partitions and evaluation protocol are used for all backbones and for all ablation studies to ensure a fair comparison.

3.6. Training and inference procedure

All models follow a two-stage training pipeline shared across backbones. In the first stage, the backbone is pre-trained solely with the Lesion-Focused Attention Loss $\mathcal{L}_{\text{LFA}}$ (Section 3.4) on the SUN training set, without attaching a classification head. This stage explicitly encourages the last-layer [CLS] token to concentrate its attention within expert-annotated lesion bounding boxes, thereby injecting spatial clinical priors into the encoder.

In the second stage, a linear classification head is added on top of the [CLS] token, and the entire model (backbone + head) is fine-tuned end-to-end on the same training set using a cross-entropy loss for the Paris classification task (2-class or 3-class, as defined in Section 3.1). The attention supervision loss $\mathcal{L}_{\text{LFA}}$ is not used during this fine-tuning stage.

Hyperparameters are kept fixed across all backbones and in both stages: we use the AdamW optimiser with an initial learning rate of 1×10^{-6}, weight decay 5×10^{-5}, batch size 32 and a maximum of 50 epochs. The learning rate is reduced on plateau using a `ReduceLROnPlateau` scheduler (factor 0.1, patience 5), and early stopping is triggered if the validation loss does not improve for 15 consecutive epochs. The same data-augmentation pipeline is applied in pretraining and fine-tuning, combining geometric and photometric transforms (horizontal/vertical flips, 90° rotations, colour jitter, Gaussian blur, elastic deformation, artificial shadows and motion blur), bounding boxes are transformed consistently with the corresponding image.

At the end of each epoch, we compute the loss and accuracy on the validation split. For each stage, the checkpoint corresponding to the lowest validation loss is retained: the best pretraining checkpoint is used to initialise the fine-tuning stage, and the best fine-tuning checkpoint is used for all test-time evaluations and attention analyses reported in Sections 4 and 5

At inference time, each frame is resized to 224 × 224 and passed once through the model without test-time augmentation. Class predictions are obtained from the softmax-normalised logits of the classification head. The last-layer `[CLS]` self-attention is extracted and processed as described in Section 3.3 to compute the attention–inside-lesion score (AttIn) and to visualise attention heatmaps.

In the ablation study, we additionally implement the Grad-CAM consistency regularisation of Lee and Cho [34], starting from the official public repository and inserting the same early-stopping and scheduling strategy described above to ensure a comparable training and inference protocol across all methods.

3.7. Evaluation metrics

The effectiveness of attention-supervised training driven by $\mathcal{L}_{\text{LFA}}$ is assessed using both conventional classification measures and metrics that directly quantify the spatial alignment of attention with lesion annotations. All metrics are computed on the held-out test split, which is never seen during training or validation, and reported values are averaged over 10 independent runs with random seeds from 0 to 9.

Classification accuracy. Overall accuracy is defined as the ratio of correct predictions to the total number of predictions:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \tag{6}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively. In the multi-class setting, the numerator corresponds to the sum of the diagonal elements of the confusion matrix and the denominator to the total number of test frames.

*Attention-inside-lesion score (*AttIn*).* For every test frame, we compute the normalised AttIn score AttIn as defined in Eq. (3). By construction, $\text{AttIn} \in [0,1]$ represents the fraction of the normalised `[CLS]` attention mass that falls inside the ground-truth lesion bounding box. In contrast to overlap measures such as the Dice coefficient, AttIn does not require binarising the attention map and is therefore less sensitive to arbitrary threshold choices. The mean and standard deviation of AttIn across the test set provide a global indicator of how faithfully the model concentrates attention within annotated lesions.

Stratified accuracy. To further assess trustworthiness, we stratify the test set according to AttIn. Accuracy (Eq. (6)) is reported separately for a low-AttIn subset ($\text{AttIn} \le 0.5$) and a high-AttIn subset ($\text{AttIn} > 0.5$). A higher accuracy in the high-AttIn subset supports the use of AttIn as an intrinsic confidence signal.

Kernel-density estimation (KDE). To characterise the relationship between spatial attention alignment and prediction correctness, we analyse the distribution of per-frame AttIn (Eq. (3)) using KDE. Separate KDEs are computed for correctly and incorrectly classified samples with seaborn kdeplot function, using a Gaussian kernel and a bandwidth adjustment of $h = 0.75$. The resulting density estimates compactly describe the empirical distributions of AttIn for correct versus incorrect predictions. Marked separation between these densities indicates that larger AttIn values are associated with higher classification reliability, providing a visual confirmation of attention–inside-lesion mass as an intrinsic confidence indicator.

Chi-squared test χ^2. A χ^2 test evaluates the dependence between the binary variable $\text{AttIn} > 0.5$ and prediction correctness. $p < 0.05$ indicates that spatial alignment and performance are not independent, reinforcing the validity of AttIn as a proxy for trust.

3.8. Statistical tests

To assess the significance of performance differences across model variants and runs, we use two-sided Welch t-tests on metrics computed over 10 runs with different random seeds. For each task and backbone, we report the Welch t-test p-values comparing attention-supervised pretraining against the corresponding non-pretrained baseline, and we consider differences statistically significant when $p < 0.05$.

4. Results

4.1. Effect of attention-supervised pretraining on accuracy and overlap

Table 2 summarises the effect of attention-supervised pretraining on both accuracy and lesion-focused attention for all backbones. Below, we discuss the 2-class and 3-class settings separately.

In the 2-class scenario, attention supervision primarily boosts lesion-focused attention while having a more heterogeneous impact on accuracy. For all backbones, AttIn increases markedly and significantly (e.g. ViT-B/16 from 0.10 to 0.34, DINOv2-L from 0.10 to 0.76; all $\Delta\text{AttIn} > 0$ with $p < 0.05$; Table 2). In terms of accuracy, the clearest benefit is observed for ViT-B/16, which improves from 68.8% to 71.9% (ΔAccuracy +3.1 percentage points, $p < 0.05$). DINOv2-B and DINOv2-L also show modest gains (+2.25 and +1.95 percentage points, respectively), although these are not statistically significant. In contrast, the two 32-pixel patch models exhibit significant accuracy drops despite large AttIn gains: ViT-B/32 decreases from 63.2% to 54.7% and ViT-L/32 from 71.2% to 61.4% (both $p < 0.05$), suggesting that enforcing tight, lesion-centred attention at very coarse spatial resolution can conflict with the classification objective.

In the more challenging 3-class setting, attention-supervised pretraining yields stronger and more consistent improvements in both accuracy and overlap. As in the 2-class case, all backbones show large and significant increases in AttIn (e.g. ViT-L/16 from 0.08 to 0.73, DINOv2-L from 0.10 to 0.86; Table 2). Crucially, accuracy gains are now more widespread and statistically robust: ViT-B/16 improves from 50.7% to 58.1% (+7.4 points, $p < 0.05$), DINOv2-B from 53.9% to 58.5% (+4.61 points, $p < 0.05$), DINOv2-L from 51.8% to 54.6% (+2.81 points, $p < 0.05$), and ViT-B/32 from 46.9% to 48.7% (+1.76 points, $p < 0.05$). ViT-L/16 and ViT-L/32 maintain accuracy within the error margins while substantially increasing AttIn. Overall, final accuracies around 58% for ViT-B/16 and DINOv2-B are achieved together with clearly lesion-focused attention distributions.

Taken together, Table 2 shows that attention-supervised pretraining reliably increases lesion-focused attention for all architectures and tends to improve or at least preserve accuracy, with more frequent and statistically significant gains in the 3-class task than in the 2-class task. The method is particularly effective for backbones with sufficient spatial granularity (16-pixel patches) DINOv2 pretraining, while ViT-L/32 in the 2-class setting remains the main exception where improved overlap comes at a clear cost in accuracy.

Table 2
Effect of attention-supervised pretraining on accuracy and lesion-focused attention. Mean ± SD test accuracy (in %) and attention–inside-lesion score AttIn are reported for each backbone with and without pretraining across multiple runs. ΔAccuracy and ΔAttIn are absolute changes (pretrained minus non-pretrained). Asterisks on ΔAccuracy and ΔAttIn indicate $p < 0.05$ for the corresponding comparison between pretrained and non-pretrained models based on tests over seed-matched runs. **Bold** marks the best and underline the second-best accuracy within each block.

	Model	Pretrain	Accuracy (%)	AttIn	ΔAccuracy (%)	ΔAttIn
2-Class	ViT-B/16	×	68.8 ± 2.0	0.101 ± 0.016		
	ViT-B/16	✓	**71.9 ± 1.9**	0.343 ± 0.112	+3.1*	+0.24*
	ViT-L/16	×	71.4 ± 5.7	0.087 ± 0.007		
	ViT-L/16	✓	68.8 ± 6.6	0.749 ± 0.021	−2.55	+0.662*
	ViT-B/32	×	63.2 ± 4.6	0.079 ± 0.017		
	ViT-B/32	✓	54.7 ± 1.6	0.230 ± 0.045	−8.56*	+0.151*
	ViT-L/32	×	71.2 ± 3.7	0.049 ± 0.004		
	ViT-L/32	✓	61.4 ± 2.7	0.410 ± 0.036	−9.87*	+0.361*
	DINOv2-B	×	65.2 ± 4.8	0.124 ± 0.020		
	DINOv2-B	✓	67.5 ± 2.7	0.770 ± 0.139	+2.25	+0.646*
	DINOv2-L	×	60.7 ± 4.0	0.102 ± 0.039		
	DINOv2-L	✓	62.6 ± 3.8	0.761 ± 0.039	+1.95	+0.659*
3-Class	ViT-B/16	×	50.7 ± 4.3	0.095 ± 0.015		
	ViT-B/16	✓	58.1 ± 1.1	0.322 ± 0.034	+7.4*	+0.23*
	ViT-L/16	×	56.9 ± 5.1	0.077 ± 0.008		
	ViT-L/16	✓	57.5 ± 3.4	0.733 ± 0.026	+0.58	+0.656*
	ViT-B/32	×	46.9 ± 2.4	0.087 ± 0.013		
	ViT-B/32	✓	48.7 ± 3.2	0.143 ± 0.061	+1.76*	+0.056*
	ViT-L/32	×	48.3 ± 4.3	0.055 ± 0.005		
	ViT-L/32	✓	49.9 ± 2.8	0.179 ± 0.048	+1.60	+0.124*
	DINOv2-B	×	53.9 ± 4.2	0.122 ± 0.019		
	DINOv2-B	✓	**58.5 ± 3.0**	0.794 ± 0.066	+4.61*	+0.673*
	DINOv2-L	×	51.8 ± 3.0	0.096 ± 0.021		
	DINOv2-L	✓	54.6 ± 2.0	0.857 ± 0.014	+2.81*	+0.762*

4.2. Relationship between attention overlap and prediction correctness

The relationship between spatial attention alignment and predictive reliability is summarised in Table 3, using a high-overlap threshold of AttIn > 0.5. For all non-pretrained models, only a very small fraction of test frames (typically around 1%) exceeds this threshold, and the mean AttIn among correctly classified samples remains low. After attention-supervised pretraining, the proportion of high-overlap frames increases dramatically, reaching 60%–80% for several large backbones in both classification tasks, and the mean AttIn for correct predictions becomes substantially higher and, in most cases, significantly larger than for incorrect ones (as indicated by the superscript * on Mean AttIn in Table 3).

In the 2-class task, attention supervision yields a clear coverage gain at high overlap: the fraction of frames with AttIn > 0.5 rises from about 1% in the baseline models to 30%–80% for ViT-L/16, ViT-L/32, DINOv2-B and DINOv2-L. For all pretrained backbones, the high-overlap subset is more accurate than the low-overlap subset, with ΔAcc consistently positive and often above 10 percentage points (e.g. +23.70 for ViT-L/16 and +10.92 for ViT-L/32). The χ^2 test confirms a statistically significant dependence between correctness and the high-overlap indicator for several models, notably ViT-L/16, ViT-B/32 and ViT-L/32, indicating that lesion-focused attention and prediction correctness are not independent in this setting.

In the 3-class task, the same qualitative picture holds, with even stronger contrasts for many architectures. After pretraining, large models such as ViT-L/16, ViT-B/32 and DINOv2-L achieve large positive gaps between high- and low-overlap subsets (ΔAcc of +33.62, +32.50 and +35.17 percentage points, respectively), together with highly significant χ^2 p-values ($<10^{-5}$). The main exception is ViT-B/16, where at the AttIn > 0.5 threshold the high-overlap subset is slightly less accurate than the low-overlap subset (ΔAcc −4.45), despite having substantially higher mean AttIn. For the remaining backbones, however, high-overlap frames are consistently more accurate, and the statistical association between overlap and correctness is particularly strong for the larger ViT and DINOv2 architectures.

Overall, these results show that attention-supervised pretraining does not merely increase the amount of attention inside the lesion; it also turns attention–lesion overlap into a meaningful confidence signal. Frames in which the model concentrates a large fraction of its attention within the annotated polyp region are, for most backbones and especially in the 3-class task, markedly more likely to be correctly classified than frames with diffuse or off-lesion attention.

4.3. KDE analysis

The KDE in Fig. 3 further characterise the relationship between overlap and correctness. In the non-pretrained setting (top row), all three backbones (ViT-L/16, ViT-L/32 and DINOv2-L) show very similar AttIn distributions for correct and incorrect predictions: both are tightly concentrated near zero, with means well below 0.1 and almost no mass beyond the threshold $\tau = 0.5$, indicating that raw attention provides little information about reliability. After pretraining with $\mathcal{L}_{\text{LFA}}$ (bottom row), ViT-L/16 and DINOv2-L exhibit a marked shift: correct predictions develop a dominant mode at high overlap values (around 0.8–1.0), while incorrect predictions remain peaked near zero, and the threshold $\tau = 0.5$ lies between these two regimes. For ViT-L/32, the separation is weaker but still visible, with the correct distribution shifted to the right and a heavier high-overlap tail compared to the incorrect one. These KDEs visually confirm the tabulated results: attention supervision turns AttIn into a much more discriminative, almost monotonic indicator of correctness for large models with sufficient spatial resolution, whereas in the non-pretrained case overlap is largely uninformative.

Table 3
Attention overlap, prediction correctness and high-overlap subset accuracy. Values are mean ± SD across seed-matched runs using the AttIn threshold 0.5. Samples AttIn > 0.5 is the percentage of test frames above threshold. Mean AttIn (correct) is computed on correctly classified frames, superscript * on Mean AttIn indicates $p < 0.05$ for the run-wise gap in AttIn between correct and incorrect predictions. Accuracy AttIn > 0.5 is accuracy on the high-overlap subset, and ΔAcc AttIn > 0.5 is (high-overlap accuracy minus low-overlap accuracy). The χ^2 column reports the mean p-value across runs for the test of dependence between correctness and the indicator AttIn > 0.5, written in scientific notation.

	Model	Pretrain	Samples $AttIn > 0.5$ (%)	Mean AttIn (correct)	Accuracy $AttIn > 0.5$ (%)	ΔAcc $AttIn > 0.5$ (%)	χ^2 p-value
2-Class	ViT-B/16	×	1.4	0.113 ± 0.016	90.6 ± 6.0	+22.07	3.60×10^{-2}
	ViT-B/16	✓	30.9	0.358* ± 0.089	77.7 ± 8.9	+6.76	2.82×10^{-1}
	ViT-L/16	×	0.7	0.096 ± 0.014	90.8 ± 6.7	+22.94	1.08×10^{-1}
	ViT-L/16	✓	79.5	0.807* ± 0.021	70.3 ± 6.2	+23.70	$< 10^{-5}$
	ViT-B/32	×	0.8	0.097 ± 0.027	94.1 ± 7.0	+31.18	1.52×10^{-1}
	ViT-B/32	✓	13.8	0.251* ± 0.055	58.9 ± 9.5	+6.17	2.70×10^{-2}
	ViT-L/32	×	0.1	0.049 ± 0.004	89.9 ± 17.9	+18.82	5.92×10^{-1}
	ViT-L/32	✓	41.6	0.433* ± 0.024	73.4 ± 3.0	+10.92	$< 10^{-5}$
	DINOv2-B	×	1.2	0.151 ± 0.030	97.1 ± 3.2	+33.69	3.00×10^{-3}
	DINOv2-B	✓	79.9	0.796* ± 0.144	67.8 ± 2.4	+10.53	7.70×10^{-2}
	DINOv2-L	×	0.8	0.126 ± 0.044	97.4 ± 3.5	+37.08	9.00×10^{-3}
	DINOv2-L	✓	78.2	0.792* ± 0.035	64.4 ± 4.6	+7.94	1.05×10^{-1}
3-Class	ViT-B/16	×	1.1	0.113 ± 0.016	90.6 ± 6.2	+22.07	3.58×10^{-2}
	ViT-B/16	✓	28.5	0.310 ± 0.033	55.0 ± 1.8	−4.45	9.05×10^{-2}
	ViT-L/16	×	1.0	0.107 ± 0.017	63.7 ± 15.5	+6.83	2.78×10^{-1}
	ViT-L/16	✓	59.3	0.695 ± 0.037	71.1 ± 3.6	+33.62	$< 10^{-5}$
	ViT-B/32	×	1.0	0.096 ± 0.022	84.9 ± 7.5	+42.78	1.00×10^{-3}
	ViT-B/32	✓	27.4	0.384 ± 0.026	73.0 ± 8.3	+32.50	$< 10^{-5}$
	ViT-L/32	×	0.1	0.056 ± 0.008	68.1 ± 24.6	+19.46	5.44×10^{-1}
	ViT-L/32	✓	9.3	0.218 ± 0.064	74.8 ± 12.4	+28.06	8.00×10^{-2}
	DINOv2-B	×	1.1	0.158 ± 0.027	94.2 ± 4.9	+40.83	3.00×10^{-3}
	DINOv2-B	✓	82.6	0.813 ± 0.067	59.0 ± 2.6	+3.44	1.76×10^{-1}
	DINOv2-L	×	1.2	0.128 ± 0.030	72.7 ± 9.8	+20.86	5.20×10^{-2}
	DINOv2-L	✓	88.2	0.932 ± 0.014	58.8 ± 2.8	+35.17	$< 10^{-5}$

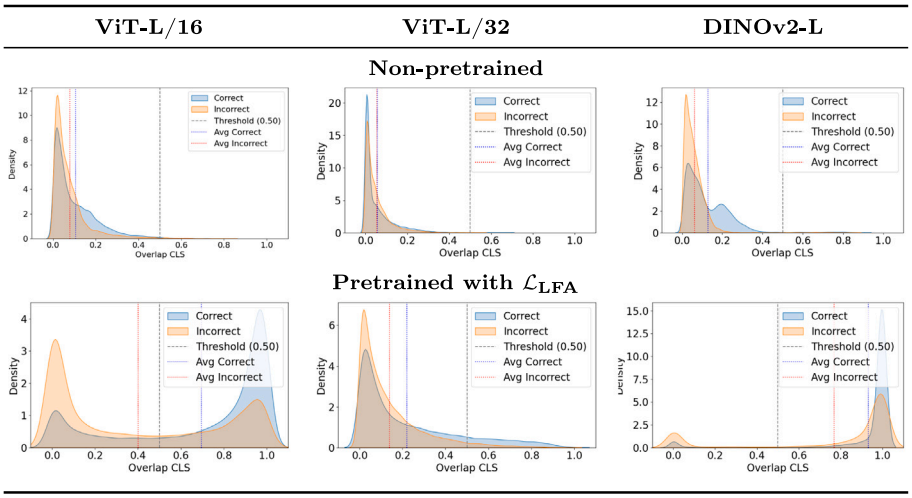


Fig. 3. Distributions of lesion-focused attention for correct and incorrect predictions. KDE of AttIn are shown for three ViT backbones, comparing correct and incorrect predictions before and after attention-supervised pretraining in the 3-class task.

4.4. Attention map analysis

Qualitative examples in Table 4 support the quantitative trends. Before pretraining, ViT-L/16 attention maps are often diffuse, with substantial focus on background structures even when the class prediction is correct. After attention-supervised pretraining, attention becomes compact and consistently centred on the polyp region: correct predictions typically display tight, lesion-conforming hotspots, and even many misclassified frames still highlight the lesion rather than unrelated areas. These examples illustrate how the proposed pretraining stabilises the spatial pattern of attention, making the intrinsic explanations more consistent with expert annotations.

4.5. Ablation study

Normalisation effect and literature supervision strategies. The ablation on ViT-B/16 (Table 5) provides a head-to-head comparison between the proposed attention-supervised pretraining and the Grad-CAM–based consistency baseline. To assess also the effect of $\mathcal{L}_1$ we have also reported the results without using it. In the 2-class task, the fully

Table 4
Qualitative comparison of [CLS] attention maps. Example attention heatmaps from ViT-L/16 without (top row) and with (bottom row) attention-supervised pretraining, stratified by prediction correctness and whether attention lies mainly inside or outside the lesion. Attention-supervised models produce more consistent lesion-centred focus across both correct and incorrect cases.

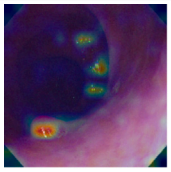
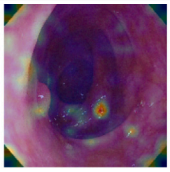
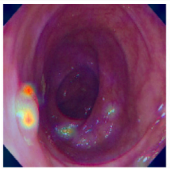
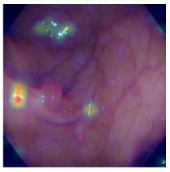
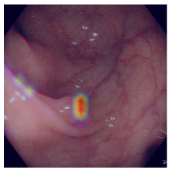
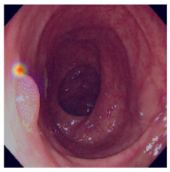
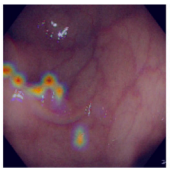
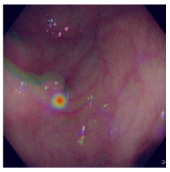
Model	Correct & Attention Inside	Correct & Attention Outside	Incorrect & Attention Inside	Incorrect & Attention Outside
Non-Pretrained (ViT-L/16)	 GT: 0-Is, Pred: 0-Is (a)	 GT: 0-Is, Pred: 0-Is (c)	 GT: 0-Is, Pred: 0-Ip (e)	 GT: 0-IIa, Pred: 0-Is (g)
Pretrained with $\mathcal{L}_{\text{LFA}}$ (ViT-L/16)	 GT: 0-IIa, Pred: 0-IIa (b)	 GT: 0-Is, Pred: 0-Is (d)	 GT: 0-IIa, Pred: 0-Is (f)	 GT: 0-IIa, Pred: 0-Is (h)

Table 5
Ablation of normalisation and Grad-CAM consistency, with high-overlap subset statistics. For ViT-B/16 in the 2-class and 3-class tasks, the table reports overall test accuracy, mean attention–inside-lesion score (AttIn), the proportion of test frames with AttIn > 0.5, and the accuracy on this high-overlap subset, under baseline training, Grad-CAM consistency [34] and the proposed $\mathcal{L}_{\text{LFA}}$ with or without $\mathcal{L}_1$ normalisation of the attention map. **Bold** and underline indicate best and second-best overall accuracies within each block. Superscript * on Accuracy or Mean AttIn indicates $p < 0.05$ for the corresponding difference from the baseline model within the same task.

Task	Pretraining method	Attn. Map $\mathcal{L}_1$ norm.	Accuracy (%)	Mean AttIn	Samples $AttIn > 0.5$ (%)	Accuracy $AttIn > 0.5$ (%)
2-class	Baseline		68.8 ± 2.0	0.101 ± 0.016	1.1	90.6 ± 6.2
	GCAM-cons. [34]		66.5 ± 4.2	0.879 ± 0.000*	87.5	66.5 ± 4.2
	$\mathcal{L}_{\text{LFA}}$ (ours)	×	68.5 ± 2.8	0.043 ± 0.007*	4.3	80.3 ± 8.1
	$\mathcal{L}_{\text{LFA}}$ (ours)	✓	**71.9 ± 1.9***	0.343 ± 0.112*	30.9	77.7 ± 8.9
3-class	Baseline		50.7 ± 4.3	0.095 ± 0.015	0.9	79.2 ± 12.9
	GCAM-cons. [34]		45.2 ± 2.4*	0.875 ± 0.000*	87.5	45.3 ± 2.4
	$\mathcal{L}_{\text{LFA}}$ (ours)	×	53.0 ± 2.1*	0.006 ± 0.001*	0.3	80.8 ± 5.6
	$\mathcal{L}_{\text{LFA}}$ (ours)	✓	**58.1 ± 1.1***	0.322 ± 0.034*	28.5	55.0 ± 1.8

normalised attention-supervised model achieves the highest accuracy (about 72%), improving over both the baseline and Grad-CAM, and increases lesion-focused overlap relative to the non-pretrained model. In the 3-class task, the same configuration again yields the best performance.

By contrast, Grad-CAM consistency substantially modifies attention but does not improve accuracy: it reduces classification performance compared with the baseline and drives overlap to extreme values, indicating that the model has learned a degenerate attention pattern rather than meaningful lesion focus. The variant without normalisation achieves modest accuracy gains but collapses overlap towards near-zero values. Taken together, these results show that attention-supervised pretraining is more effective than Grad-CAM–based regularisation for jointly improving accuracy and spatial alignment, and explicit attention normalisation is essential to avoid degenerate solutions and to obtain interpretable overlap scores.

Threshold sensitivity analysis. Table 6 shows how the reliability signal changes with the AttIn threshold τ. In the 2-class task, increasing τ for the attention-supervised model selects progressively smaller high-overlap subsets (from 41.9% at $\tau = 0.3$ to 14.0% at $\tau = 0.9$), while the accuracy gain ΔAccuracy remains positive and the χ^2 p-value decreases, reaching 7.60×10^{-4} at $\tau = 0.9$, i.e. the strongest dependence between overlap and correctness.

In the 3-class task, the pretrained ViT-B/16 does not initially show a reliability gain for high-overlap frames: at $\tau = 0.3$ and $\tau = 0.5$, ΔAccuracy is slightly negative despite moderate coverage (43.9% and 28.5%). However, at the more conservative threshold $\tau = 0.9$, the high-overlap subset becomes both more accurate (+8.3 percentage points) and significantly associated with correctness ($p = 4.30 \times 10^{-2}$). Overall, this indicates that, after attention-supervised pretraining, overlap can be used as a tunable intrinsic confidence criterion, with higher thresholds trading coverage for a stronger accuracy and χ^2 signal.

5. Discussion

5.1. Model accuracy

Across backbones and both Paris classification tasks, attention-supervised pretraining with $\mathcal{L}_{\text{LFA}}$ generally preserves or improves accuracy while markedly increasing lesion-focused attention (Table 2). In the three-class setting, the effect is particularly consistent: all backbones either match or outperform their non-pretrained counterparts, with the strongest gains observed for ViT-B/16 and DINOv2-B. This suggests that the additional spatial prior provided by $\mathcal{L}_{\text{LFA}}$ is especially beneficial in the more challenging multi-class morphology scenario. In the two-class task, the picture is more heterogeneous: architectures with finer spatial granularity (16-pixel patches) and the DINOv2 backbones still profit from attention supervision, whereas models with 32-pixel patches show a trade-off in which lesion-focused attention increases but accuracy can decline (Table 2). This pattern indicates

Table 6
Sensitivity of high-overlap subsets to the AttIn **threshold.** For ViT-B/16 in the 2-class and 3-class tasks, the table shows, for several thresholds τ, the proportion of test frames with AttIn $> \tau$, the accuracy gain ΔAccuracy for AttIn $> \tau$ (in percentage points) defined as the difference between high- and low-overlap subset accuracies, and the χ^2 p-value for dependence between correctness and the high-overlap indicator, comparing non-pretrained and attention-pretrained models.

	τ	Non-pretrained			Pretrained with $\mathcal{L}_{\text{LFA}}$		
		Samples (AttIn $> \tau$) (%)	ΔAccuracy (AttIn $> \tau$) (%)	χ^2 p-value	Samples ($AttIn > \tau$) (%)	ΔAccuracy (AttIn $> \tau$) (%)	χ^2 p-value
2-class	0.3	7.3	+16.0	2.40e−02	41.9	+6.4	1.32e−01
	0.5	1.1	+22.1	3.58e−02	30.9	+6.8	2.82e−01
	0.9	0.0	–	–	14.0	+13.9	7.60e−04
3-class	0.3	7.0	+23.5	1.35e−01	43.9	−5.1	7.50e−02
	0.5	0.9	+28.8	1.80e−01	28.5	−4.5	9.05e−02
	0.9	0.0	–	–	3.5	+8.3	4.30e−02

that forcing highly localised attention on a very coarse token grid may conflict with the classification objective, while encoders with sufficient spatial resolution can better exploit lesion-centred guidance. The comparatively smaller gains in the two-class scenario also suggest that attention-guided pretraining is most helpful when the decision boundary is more complex, as in the three-class task.

The ablation on ViT-B/16 clarifies the role of attention normalisation (Table 5). When $\mathcal{L}_{\text{LFA}}$ is applied *without* $\mathcal{L}_1$ normalisation, accuracy improves only modestly over the baseline and the resulting attention maps exhibit almost vanishing AttIn values. This indicates that, in the absence of $\mathcal{L}_1$, the model can minimise the loss by globally shrinking attention magnitudes rather than by reallocating mass into the lesion region. Including $\mathcal{L}_1$ explicitly regularises the scale of the attention maps, forcing them to retain a meaningful distribution of mass and thereby turning $\mathcal{L}_{\text{LFA}}$ into a genuine spatial reweighting signal rather than a simple attenuation term. The fully normalised variant consequently achieves both higher accuracy and substantially more informative overlap statistics (Table 5).

5.2. Explainability

From an explainability perspective, $\mathcal{L}_{\text{LFA}}$ systematically reshapes intrinsic [CLS] attention. After attention-supervised pretraining, all backbones allocate a larger fraction of their normalised attention mass inside the lesion bounding box, as reflected by higher AttIn values in Table 2. This effect is particularly clear for larger ViT and DINO models and in the three-class task, where the combination of a more challenging decision boundary and explicit spatial guidance appears to be especially effective. Crucially, these gains in alignment are obtained while preserving or improving classification accuracy (Table 2), indicating that the models are not merely reshuffling attention but are learning to base their decisions more consistently on clinically relevant image regions.

The link between overlap and correctness is also strengthened. After pretraining, the proportion of test frames with high overlap increases markedly for most backbones, and correctly classified frames exhibit higher AttIn than misclassified ones (Table 3). High-overlap subsets are generally more accurate than their low-overlap counterparts, with the effect again most pronounced for the larger models and in the three-class setting. Even when the overall accuracy gain is modest, the high-overlap subset tends to be substantially more reliable (Table 3). This pattern suggests that the model is not only encouraged to attend to the lesion, but also to associate lesion-centred attention with correct predictions, i.e. to trust frames in which its focus coincides with the annotated region.

The threshold-sensitivity analysis further supports the interpretation of overlap as a tunable intrinsic confidence signal (Table 6). For the attention-supervised ViT-B/16 in the two-class task, increasing the AttIn threshold gradually reduces the proportion of frames classified as high-overlap, while amplifying both the accuracy gap between high- and low-overlap subsets and the statistical dependence between overlap and correctness. In the three-class task, moderate thresholds retain broad coverage but yield limited reliability gains, whereas more conservative thresholds isolate a smaller subset of frames that is both more accurate and more strongly associated with correctness (Table 6). Taken together, these observations indicate that attention–lesion overlap can be used to trade coverage for reliability in a controlled way, providing a principled knob to derive conservative, high-trust subsets from an intrinsic model signal.

KDE of AttIn offer a complementary, distributional view of this behaviour (Fig. 3). Before pretraining, the overlap distributions for correct and incorrect predictions largely coincide and are concentrated near low values, suggesting that raw attention carries little information about reliability. After training with $\mathcal{L}_{\text{LFA}}$, the distributions separate: for the examined backbones in the three-class task, correct predictions shift towards higher overlap regions, while incorrect predictions remain concentrated at low overlap. This separation visually confirms that, once attention is supervised, the models tend to focus strongly on the lesion precisely when they are more likely to be correct.

Qualitative examples of ViT-L/16 attention maps reinforce this interpretation (Table 4). In the non-pretrained model, attention is often diffuse and spills over to background structures, even when the classification is correct. After attention-supervised pretraining, the maps become compact and lesion-centred, and many misclassified frames still highlight the polyp rather than unrelated areas. This indicates a genuine change in model behaviour: the model not only improves its classification performance but also learns to ground its decisions in the anatomically relevant region of the image, making its internal reasoning more consistent with expert annotations.

5.3. Advantages of the $\mathcal{L}_{LFA}$ framework

The $\mathcal{L}_{\text{LFA}}$ framework offers several advantages, supported by both quantitative and qualitative results. First, it leverages expert lesion localisation to directly supervise intrinsic self-attention, rather than relying on post-hoc saliency maps or auxiliary localisation heads. This direct guidance is reflected in the consistent increase in AttIn across backbones and in the emergence of high-overlap subsets that are more accurate and statistically associated with correctness (Tables 2 and 3). Attention maps thus become not only more lesion-centred, but also more tightly linked to prediction reliability, enabling overlap-based confidence stratification.

Second, the two-stage design is conceptually simple and empirically effective. Pretraining under $\mathcal{L}_{\text{LFA}}$ shapes the backbone to focus its attention on the lesion before any classification objective is applied, and fine-tuning then uses a standard cross-entropy loss. The ablation study on ViT-B/16 shows that, with $\mathcal{L}_1$ normalisation, this decoupled strategy yields the best trade-off between accuracy and overlap: without normalisation, the model can reduce the loss by shrinking

attention magnitudes, whereas with normalisation it is forced to reallocate mass into the lesion, producing well-scaled AttIn values that support meaningful thresholding (Table 5).

Third, compared with Grad-CAM consistency, $\mathcal{L}_{\mathrm{LFA}}$ provides a more suitable mechanism for steering transformer-based models. In our experiments, the Grad-CAM consistency baseline consistently reduces accuracy relative to the non-pretrained model and drives overlap towards extreme or saturated values (Table 5). This behaviour suggests that constraining a post-hoc saliency signal can lead the network to adopt degenerate explanation patterns without improving its underlying representations. By contrast, $\mathcal{L}_{\mathrm{LFA}}$ acts directly on intrinsic [CLS] self-attention, which is the mechanism that actually mediates information flow in ViTs, and was designed with explicit normalisation to avoid trivial solutions. The fact that $\mathcal{L}_{\mathrm{LFA}}$-pretrained models achieve higher accuracy and more informative overlap statistics than both the baseline and Grad-CAM consistency indicates that supervising attention at the level of the backbone, rather than regularising external explanations originally developed for CNNs, is a more effective and robust strategy for improving both performance and interpretability in Paris classification.

5.4. Limitations

The main limitation of $\mathcal{L}_{\mathrm{LFA}}$ is its dependence on precise lesion localisation. The loss uses expert bounding boxes as direct supervision for attention, so noisy, inconsistent or incomplete annotations can misguide the training signal and weaken the improvements observed in accuracy and overlap (Tables 2–6). Producing high-quality bounding boxes is labour-intensive and may not be feasible at scale or in centres with heterogeneous annotation practices.

A second limitation is the scope of the current evaluation. All experiments are conducted on a single public dataset of colorectal lesions acquired with high-definition white-light colonoscopy, and all quantitative findings refer to this setting (Tables 2–6). The generalisability of $\mathcal{L}_{\mathrm{LFA}}$ to other organs, imaging modalities, lesion types and labelling protocols remains to be explored.

Finally, the two-stage training pipeline introduces additional computational cost compared with a single end-to-end classification run. However, explanation-regularised baselines such as Grad-CAM consistency are even more demanding and, in our experiments, deliver weaker improvements in both accuracy and attention alignment (Table 5). Future work should investigate more efficient pretraining schedules, weaker or cheaper forms of spatial supervision, and cross-dataset evaluations to better understand the trade-offs between annotation effort, computational budget and the gains in accuracy and clinically meaningful attention that $\mathcal{L}_{\mathrm{LFA}}$ provides.

6. Conclusion

This work tackled two coupled limitations of Vision Transformers for Paris morphology classification on colonoscopy images: substantial inter-observer variability in human labels and diffuse, often off-lesion attention patterns that hinder interpretability. We introduced a $\mathcal{L}_{\mathrm{LFA}}$ that directly supervises last-layer [CLS] attention with expert polyp bounding boxes during a dedicated pretraining stage, before standard cross-entropy fine-tuning.

Across six ViT architectures and both 2-class (0-I vs. 0-II) and 3-class (0-Ip, 0-Is, 0-IIa) tasks on the public SUN dataset, attention-supervised pretraining with $\mathcal{L}_{\mathrm{LFA}}$ consistently increased lesion-focused attention and, for most backbones, improved accuracy. In the 3-class setting, $\mathcal{L}_{\mathrm{LFA}}$ yielded gains of up to 7 percentage points for ViT-B/16 and 5 percentage points for DINOv2-B, while driving mean attention–inside-lesion scores from low values around 0.1 to clearly lesion-centred regimes. In an ablation on ViT-B/16, $\mathcal{L}_{\mathrm{LFA}}$ also outperformed a Grad-CAM consistency baseline, indicating that directly supervising intrinsic attention is more effective than regularising post-hoc saliency for transformer-based Paris classification.

Importantly, $\mathcal{L}_{\mathrm{LFA}}$ not only increases attention–lesion overlap but also strengthens its link with prediction correctness: in most cases high-overlap subsets are more accurate, and the distributions of overlap for correct versus incorrect predictions become clearly separated. This turns lesion-focused attention into a practical intrinsic confidence cue, suggesting that compact, polyp-centred attention maps can help endoscopists gauge when model predictions are more trustworthy. Future work will explore weaker spatial supervision, larger and more diverse datasets, and prospective reader studies to assess how $\mathcal{L}_{\mathrm{LFA}}$-supervised models integrate into real-time colonoscopy workflows.

CRediT authorship contribution statement

Luca Carlini: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Luca Di Stefano:** Investigation, Conceptualization. **Chiara Lena:** Supervision, Investigation, Conceptualization. **Davide Massimi:** Supervision. **Tommy Rizkala:** Supervision. **Cesare Hassan:** Supervision. **Elena De Momi:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization.

Ethics statement

This study was conducted using the publicly available SUN dataset, which contains de-identified and anonymised endoscopic images. No new data involving human subjects were collected for this research. Therefore, ethical approval and informed consent were not required.

Declaration of competing interest

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.

Acknowledgements

Luca Carlini is supported by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

Chiara Lena is supported by Multilayered Urban Sustainability Action (MUSA) project (ECS00000037), funded by the European Union – NextGenerationEU, under the National Recovery and Resilience Plan (NRRP).

Cesare Hassan is supported by the European Commission (Horizon Europe 101057099). The Associazione Italiana per la Ricerca sul Cancro (AIRC): IG 2022 – ID. 27843 project/(AIRC) IG 2023 – ID. 29220 project. And Bando PNRR-MCNT2-2023-12377041.

Code availability

The source code and the dataset splits used in this study are publicly available at https://github.com/LucaCarlini/SUNDatasetPretraining. The repository contains the implementation of the proposed loss, the training and evaluation scripts for all considered architectures, and the exact SUNf dataset splits used for the reported experiments.

References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: Cancer J. Clin. 71 (3) (2021) 209–249.

[2] M.B. Nierengarten, Colonoscopy remains the gold standard for screening despite recent tarnish: Although a recent study seemed to indicate that colonoscopies are not as effective as once thought at detecting colorectal cancer, a closer look at the study clears the confusion, Cancer (0008543X) 129 (3) (2023).

[3] N.M. Mansour, Artificial intelligence in colonoscopy, Curr. Gastroenterol. Rep. 25 (6) (2023) 122–129.

[4] C. Hassan, M. Povero, L. Pradelli, M. Spadaccini, A. Repici, Cost-utility analysis of real-time artificial intelligence-assisted colonoscopy in Italy, Endosc. Int. Open 11 (11) (2023) E1046–E1055.

[5] D.K. Rex, C. Hassan, M.J. Bourke, The colonoscopist's guide to the vocabulary of colorectal neoplasia: histology, morphology, and management, Gastrointest Endosc. 86 (2) (2017) 253–263.

[6] S.C. Van Doorn, Y. Hazewinkel, J.E. East, M.E. Van Leerdam, A. Rastogi, M. Pellisé, S. Sanduleanu-Dascalescu, B.A. Bastiaansen, P. Fockens, E. Dekker, Polyp morphology: an interobserver evaluation for the Paris classification among international experts, Off. J. Am. Coll. Gastroenterology 110 (1) (2015) 180–187.

[7] L.J. Smits, E. Vink-Börger, G. van Lijnschoten, I. Focke-Snieders, R.S. van der Post, J.B. Tuynman, N.C. van Grieken, I.D. Nagtegaal, Diagnostic variability in the histopathological assessment of advanced colorectal adenomas and early colorectal cancer in a screening population, Histopathology 80 (5) (2022) 790–798.

[8] A. Krenzer, S. Heil, D. Fitting, S. Matti, W.G. Zoller, A. Hann, F. Puppe, Automated classification of polyps using deep learning architectures and few-shot learning, BMC Med. Imaging 23 (1) (2023) 59.

[9] S. Aburass, O. Dorgham, J. Al Shaqsi, M. Abu Rumman, O. Al-Kadi, Vision transformers in medical imaging: a comprehensive review of advancements and applications across multiple diseases, J. Imaging Inform. Med. (2025) 1–44.

[10] R. Di Bidino, S. Daugbjerg, S.C. Papavero, I.H. Haraldsen, A. Cicchetti, D. Sacchini, Health technology assessment framework for artificial intelligence-based technologies, Int. J. Technol. Assess. Health Care 40 (1) (2024) e61.

[11] M. Misawa, S.-e. Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida, H. Itoh, M. Oda, K. Mori, Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video), Gastrointest Endosc. 93 (4) (2021) 960–967.e3, http://dx.doi.org/10.1016/j.gie.2020.07.060.

[12] H. Itoh, M. Misawa, Y. Mori, M. Oda, S.-E. Kudo, K. Mori, SUN colonoscopy video database, 2020.

[13] M. Walmer, S. Suri, K. Gupta, A. Shrivastava, Teaching matters: Investigating the role of supervision in vision transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7486–7496.

[14] M. Chung, J.B. Won, G. Kim, Y. Kim, U. Ozbulak, Evaluating visual explanations of attention maps for transformer-based medical imaging, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 110–120.

[15] Y. Mori, E.H. Jin, D. Lee, Enhancing artificial intelligence-doctor collaboration for computer-aided diagnosis in colonoscopy through improved digital literacy, Dig. Liver Dis. 56 (7) (2024) 1140–1143.

[16] I.D. Mienye, G. Obaido, N. Jere, E. Mienye, K. Aruleba, I.D. Emmanuel, B. Ogbuokiri, A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges, Informatics Med. Unlocked 51 (2024) 101587.

[17] W. Samek, G. Montavon, S. Lapuschkin, C.J. Anders, K.-R. Müller, Explaining deep neural networks and beyond: A review of methods and applications, Proc. IEEE 109 (3) (2021) 247–278.

[18] K. Borys, Y.A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C.M. Friedrich, F. Nensa, Explainable AI in medical imaging: An overview for clinical practitioners–Beyond saliency-based XAI approaches, Eur. J. Radiol. 162 (2023) 110786.

[19] B.H. Van der Velden, H.J. Kuijf, K.G. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022) 102470.

[20] H. Baniecki, M. Chrabaszcz, A. Holzinger, B. Pfeifer, A. Saranti, P. Biecek, Be careful when evaluating explanations regarding ground truth, 2023, arXiv preprint arXiv:2311.04813.

[21] S. Ali, F. Akhlaq, A.S. Imran, Z. Kastrati, S.M. Daudpota, M. Moosa, The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review, Comput. Biol. Med. 166 (2023) 107555.

[22] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, U.R. Acharya, Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022), Comput. Methods Programs Biomed. 226 (2022) 107161.

[23] C. Hassan, T. Rizkala, Y. Mori, M. Spadaccini, M. Misawa, G. Antonelli, E. Rondonotti, E. Dekker, B.B. Houwen, O. Pech, et al., Computer-aided diagnosis for the resect-and-discard strategy for colorectal polyps: a systematic review and meta-analysis, Lancet Gastroenterol. Hepatol. 9 (11) (2024) 1010–1019.

[24] Y. Komeda, H. Handa, T. Watanabe, T. Nomura, M. Kitahashi, T. Sakurai, A. Okamoto, T. Minami, M. Kono, T. Arizumi, et al., Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience, Oncology 93 (Suppl. 1) (2017) 30–34.

[25] M.F. Byrne, N. Chapados, F. Soudan, C. Oertel, M.L. Pérez, R. Kelly, N. Iqbal, F. Chandelier, D.K. Rex, Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model, Gut 68 (1) (2019) 94–100.

[26] S. Tanwar, P.M. Goel, P. Johri, M.J. Divan, Classification of benign and malignant colorectal polyps using pit pattern classification, in: Proceedings of the 4th International Conference: Innovative Advancement in Engineering & Technology, IAET, 2020.

[27] T. Ozawa, S. Ishihara, M. Fujishiro, Y. Kumagai, S. Shichijo, T. Tada, Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks, Ther. Adv. Gastroenterol. 13 (2020) 1756284820910659.

[28] R. Zhang, Y. Zheng, T.W.C. Mak, R. Yu, S.H. Wong, J.Y. Lau, C.C. Poon, Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain, IEEE J. Biomed. Health Inform. 21 (1) (2016) 41–47.

[29] C.-M. Lo, Y.-H. Yeh, J.-H. Tang, C.-C. Chang, H.-J. Yeh, Rapid polyp classification in colonoscopy using textural and convolutional features, in: Healthcare, 10 (8) (2022) 1494.

[30] A. Bour, C. Castillo-Olea, B. Garcia-Zapirain, S. Zahia, Automatic colon polyp classification using convolutional neural network: a case study at basque country, in: 2019 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT, IEEE, 2019, pp. 1–5.

[31] A.A. Ismail, H. Corrada Bravo, S. Feizi, Improving deep learning interpretability by saliency guided training, 2021.

[32] S. Ben Itzhak, N. Kiryati, O. Portnoy, A. Mayer, Localization-guided supervision for robust medical image classification by vision transformers, in: European Conference on Computer Vision, Springer, 2024, pp. 118–133.

[33] Y. Yu, X. Wu, P. Yu, Q. Wan, Y. Dan, Y. Xiao, Q. Wang, Location-guided lesions representation learning via image generation for assessing plant leaf diseases severity, Plant Phenom. 7 (2) (2025) 100058, http://dx.doi.org/10.1016/j.plaphe.2025.100058, URL https://www.sciencedirect.com/science/article/pii/S2643651525000640.

[34] J. Lee, S. Cho, Semi-supervised image classification with grad-CAM consistency, 2021, arXiv preprint arXiv:2108.13673.

[35] S. Zhu, Y. Zhang, Y. Feng, GW-net: An efficient grad-CAM consistency neural network with weakening of random erasing features for semi-supervised person re-identification, Image Vis. Comput. 137 (2023) 104790.

[36] T. Xiao, C.J. Reed, X. Wang, K. Keutzer, T. Darrell, Region similarity representation learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10539–10548.

[37] Y. Zhou, X. Gao, Z. Chen, H. Huang, Attention distillation: A unified approach to visual characteristics transfer, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 18270–18280.

[38] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, 2023, arXiv preprint arXiv:2304.07193.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[40] J.M. Johnson, T.M. Khoshgoftaar, Survey on deep learning with class imbalance, J. Big Data 6 (1) (2019) 1–54.