

Attention to detail: inter-resolution knowledge distillation

Rocío del Amor*, Julio Silva-Rodríguez†, Adrián Colomer* ‡, Valery Naranjo*

**Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano,*

HUMAN-tech, Universitat Politècnica de València, Valencia, Spain

Email: madeam2@upv.es

†*ÉTS Montréal, Montréal, Québec, Canada*

‡*ValgrAI- Valencian Graduate School and Research Network for Artificial Intelligence*

Abstract—The development of computer vision solutions for gigapixel images in digital pathology is hampered by significant computational limitations due to the large size of whole slide images. In particular, digitizing biopsies at high resolutions is a time-consuming process, which is necessary due to the worsening results from the decrease in image detail. To alleviate this issue, recent literature has proposed using knowledge distillation to enhance the model performance at reduced image resolutions. In particular, soft labels and features extracted at the highest magnification level are distilled into a model that takes lower-magnification images as input. However, this approach fails to transfer knowledge about the most discriminative image regions in the classification process, which may be lost when the resolution is decreased. In this work, we propose to distill this information by incorporating attention maps during training. In particular, our formulation leverages saliency maps of the target class via grad-CAMs, which guides the lower-resolution Student model to match the Teacher distribution by minimizing the L_2 -distance between them. Comprehensive experiments on prostate histology image grading demonstrate that the proposed approach substantially improves the model performance across different image resolutions compared to previous literature. The project code is available on https://github.com/cvblab/kd_resolution.

Index Terms—Knowledge distillation, Attention constraints, Inter-resolution, Histology image.

I. INTRODUCTION

Computer vision methods using deep learning have reached remarkable results in a wide range of applications, including medical image analysis. This is the case even in challenging fields such as digital pathology, where digitized biopsies take the form of whole-slide images (WSI) consisting of hundreds of thousands of pixels. In addition, the relevant areas may be contained in a small region of the image. Different successful applications of deep learning in digital pathology include global tasks such as biopsy level cancer detection [1], tissue segmentation [2] or small structures detection, such as mitotic figures [3]. In real-world scenarios, the highest image augmentations might not be available, and hardware and time constraints can present challenges to deploying these models. These limitations emphasize the need for novel and efficient solutions for implementing computer-aided systems into clinical practice. In this line, recent work has explored the feasibility of using well-known knowledge distillation formulations to reduce the required image resolution during inference [4]. However, the vanilla knowledge distillation and

feature-matching terms used in these studies do not allow for transferring relevant and discriminative regions utilized in high-resolution images. Based on these observations, we propose an attention-aware formulation to reduce the required image resolution during model deployment. The key contributions of our work can be summarized as follows:

- A novel attention-constrained formulation for inter-resolution knowledge distillation.
- We propose to train a Student model which matches the attention maps produced by a Teacher model trained with higher-resolution images by minimizing the L_2 -distance.
- In particular, we propose to transfer only *strictly-positive* gradients in the proposed AT^+ term.
- The method is validated in the context of prostate histology image grading. Using the proposed term, the model archives competitive results while requiring $8\times$ fewer augmentations during deployment.

II. RELATED WORK

Constrained classification: Constrained learning aims to regularize the training of deep learning models for image classification tasks to produce a solution that satisfies a given condition. Through this condition, the model can incorporate additional prior knowledge of the task. The main core of literature in this field is the introduction of constraints at the pixel-level response of the model, which is achieved through the regularization of attention maps. It is worth mentioning that, despite the recent widespread use of attention for the trainable blocks of popular Transformers encoders [5], we refer to attention as the saliency class-activation maps produced by the image-level classifier when applied at the pixel-level, in the form of semantic segmentation maps or processed response maps such as Grad-CAMs [6]. In this fashion, weakly supervised segmentation models can introduce object proportion constraints as priors [7]–[9], or unsupervised anomaly detection models can be forced to focus on the whole image [10]–[12]. Also, self-supervised training methods can leverage the most discriminant regions to enhance feature learning [13], or attention constraints can be used to improve fine-grained image recognition [14].

Knowledge distillation: Knowledge distillation [15] is a field of machine learning that explores the use of deep learn-

ing solutions in creating efficient models. The most popular scenario aims to reduce the model size for deployment in systems with limited resources, so-called model compression. The main core of the literature involves using a pre-trained Teacher model, trained on large-capacity settings, to transfer an optimal solution to a low-capacity Student model. Vanilla knowledge distillation (KD) [15] transfers the softmax scores of the Teacher to the Student, which enhances inter-class relations. Other popular terms match intermediate feature representations (FM) [16] between both models via an l_2 -loss. Recent works [17] have improved the base feature representation matching by indirectly using the Teacher classifier through Student features by a softmax regression (SR) of the produced logits. However, these terms do not ensure that the Student model focuses on the same image regions for classification. To alleviate this issue, [18] introduced an attention transfer loss (AT), which forces the Student model to match the attention maps produced by the Teacher. Concretely, the gradients obtained for each spatial feature representation (before flattening or global pooling) concerning the logits output by the model are used as a proxy for attention distillation. Nevertheless, as later pointed out in [6], gradients alone might produce suboptimal results for highlighting the most relevant patterns for the task at hand. In this work, we refine the AT term to (i) weight both feature representation and gradients, and (ii) use only *strictly positive* gradients, based on Grad-CAMs for attention generation [6].

Besides model compression, knowledge distillation has also been successfully applied to other computer vision fields such as semi-supervised learning [19] or multi-modal to mono-modal segmentation [20]. This work focuses on applying knowledge distillation to enable the efficient use of deep learning models at lower image resolutions, which we refer to as inter-resolution knowledge distillation [4]. Despite its importance in high-demanding image resolution fields such as digital pathology, there is limited literature in this area, and only basic vanilla KD or feature-matching distillation has been validated [4].

III. METHODS

An overview of our proposed method is depicted in Figure 1. In the following, we describe the problem formulation and each of the proposed components.

Preliminaries: In the context of image classification, we denote a neural networks classifier as $\theta = \{\theta_f, \theta_c\}$. The model is composed of a convolutional feature extractor θ_f , which processes input images x to extract a compressed pixel-level feature representation, $f_\Omega \in \mathbb{R}^{C \times \Omega}$, such that C represents the manifold dimensionality, and Ω is the spatial size. The classifier θ_c then utilizes the global-average pooling operation on the feature representation to output softmax scores \hat{y}_k for the K target categories. In a standard image classification scenario, the model is optimized to minimize the cross-entropy loss between predicted scores and labels y_k as follows:

$$\mathcal{L}_{CE} = -\frac{1}{K} \sum_{k=1}^K y_k \log(\hat{y}_k) \quad (1)$$

A. Inter-resolution knowledge distillation

Let us denote a Teacher model as θ^t , which is trained using standard cross-entropy, as shown in Eq. 1, on high-resolution input images (x). The objective of inter-resolution knowledge distillation is to train a Student model θ^s , which takes distorted input images at a lower resolution (x^*) to take into account the rich information contained in the frozen Teacher model to improve its performance. Vanilla knowledge distillation [15] distills softmax outputs from Teacher model to incorporate image-specific inter-class dependencies by cross entropy such as:

$$\mathcal{L}_{KD} = -\frac{1}{K} \sum_{k=1}^K \hat{y}_k^t \log(\hat{y}_k^s) \quad (2)$$

Since this term only provides global information, the feature matching term in [16] allows the distilling of information regarding the feature representation by minimizing the l_2 -distance of both embeddings. Applied before the global average pooling, this matching is applied spatially such that:

$$\mathcal{L}_{FM} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \|f_i^t - f_i^s\|^2 \quad (3)$$

B. Attention matching

The \mathcal{L}_{FM} term does not leverage the most relevant regions for the target class classification. This limitation was addressed in [18] by matching attention maps obtained solely by the gradients of each spatial feature representation concerning the target class. Nevertheless, later studies on attention generation [6] suggested using only strictly positive gradients, weighted according to the feature space, to obtain refined attention maps. Thus, we compute attention maps using grad-CAMs [6], defined as $a_{\Omega,k} = ReLU(\sum_c \alpha_{c,k} f_{\Omega,c})$, where α_c represents the class-wise generated gradients and are defined as $\alpha_{c,k} = \frac{\partial f_c}{\partial \hat{y}_k^s}$, using the logits before the softmax activation as the target. Then, the attention information is distilled into the Student training using the proposed AT^+ term as follows:

$$\mathcal{L}_{AT^+} = \frac{1}{K} \sum_k \frac{1}{|\Omega|} \sum_{i \in \Omega} \|a_{i,k}^t - a_{i,k}^s\|^2 \quad (4)$$

Based on the empirical observations, which are detailed in the experimental section, we include in our formulation the \mathcal{L}_{FM} and \mathcal{L}_{AT^+} terms, and we propose to train the Student using the following global criteria:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha_{FM} \mathcal{L}_{FM} + \alpha_{AT^+} \mathcal{L}_{AT^+} \quad (5)$$

$\alpha \in \mathbb{R}^+$ is the relative weight of each knowledge distillation term, which weights its importance with respect the standard cross entropy loss in Eq. 1.

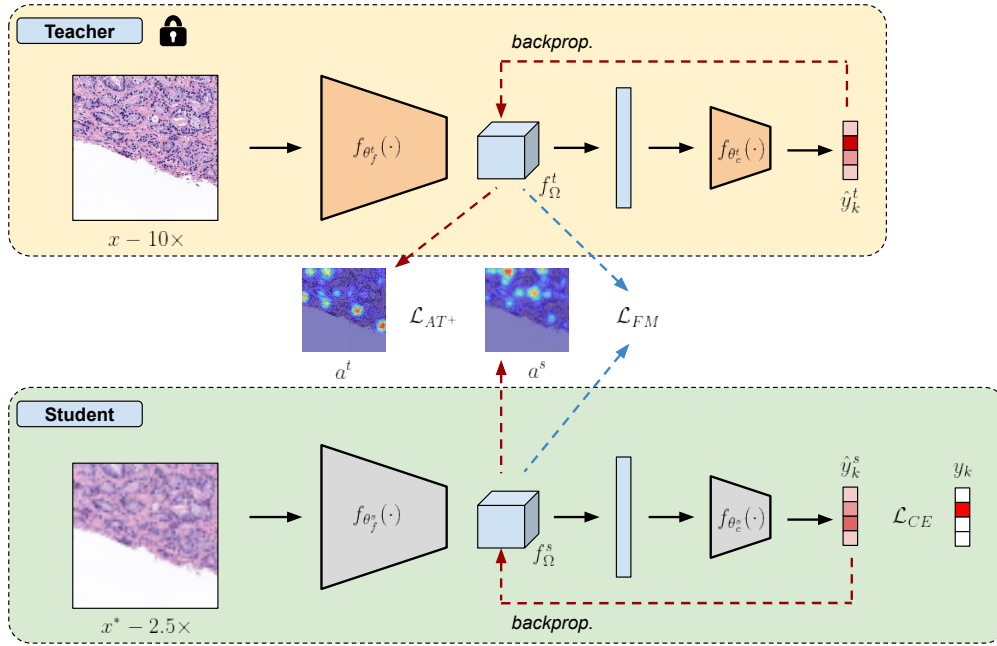


Fig. 1. **Method overview.** In the context of inter-resolution knowledge distillation, a Teacher model is trained using high-resolution images by optimizing Eq.1. To enable the deployment of efficient models that can operate at low resolutions, we train a Student model by transferring the information from the frozen Teacher. We use the well-known feature-matching distillation (Eq.3) and propose a novel attention-matching term, AT^+ (Eq.4), which distills spatial information of relevant regions in the image by using strictly positive gradient weighting for attention generation (Eq.4). Both terms are combined with the standard cross-entropy loss (Eq. 5) for the optimization of the Student model.

IV. EXPERIMENTS

A. Experimental setting

Datasets: The methods described in this work are validated in patch-level cancerous histology image grading. In particular, we use SICAPv2 [21], a public prostate histology image dataset for patch-level Gleason grading. The dataset includes 10,340 tissue patches of size 512 pixels at $10\times$ magnification. According to the presence of cancerous tissue and its severity, images have been labeled by expert pathologists using the following labels: non-cancerous tissue, Gleason grade 3, 4, and 5. The dataset presents class-balanced training, validation, and testing splits, divided at the patient level.

Implementation Details: First, Teacher model is trained using images at $10\times$ magnification via standard cross-entropy loss defined in Eq. 1. Then, the Student model is trained for different lower-magnifications inputs by optimizing the knowledge distillation criteria in Eq. 2, using the Student with frozen weights. We use images at $5\times$, $2.5\times$, and $1.25\times$ magnification as input. Low-magnification inputs are artificially created from the original $10\times$ magnification images by sequential resampling and bilinear interpolating. Teacher and Student models are initialized with the same feature extractor, VGG16 pre-trained on Imagenet. The models are trained during 20 epochs using ADAM optimizer, with a learning rate of $1e-4$ and a batch size of 32 images. Data augmentation is incorporated in random image rotations, color jitter, and random affine transforms. During training, we balance the class distribution of the training samples using proportional sampling. We monitor the performance of the model on the validation set throughout

training and select the best-performing model for testing. To determine the relative weights of each knowledge distillation term, we follow the same validation procedure with different values of $\alpha = \{0, 0.01, 0.1, 1, 10, 100\}$.

Baselines: To validate the goodness of the proposed method, we focus on relevant previous methods of knowledge distillation. Concretely, we use the vanilla knowledge distillation over softmax outputs (KD) [15], and its optimization in combination with the feature matching term (KD + FM) [4]. Also, we include softmax regression (FM + SR) by matching the logits produced by the features of Teacher and Student models through the frozen Teacher, as described in [17]. It is worth mentioning that only KD [15] and KD + FM [4] have been previously proposed in the context of inter-resolution knowledge distillation. Still, we include FM + SR [17], which obtains leading results for model compression, the core of the knowledge distillation literature. Also, we use the Student models trained at different resolutions without any knowledge distillation as a baseline. All models are trained using the same hyperparameter setting described above for the proposed method, which showed consistent performance.

Evaluation Metrics: We use standard metrics used in previous literature for disease grading in the medical context. In particular, we use the accuracy (Acc) between predicted and reference labels and Cohen’s quadratic kappa (κ), which considers inter-rater agreement in the context of ordered categories classification. To ensure reproducibility and account for random events in weight initialization, we repeat all experiments using three fixed seeds and average the results across repetitions.

B. Results

Comparison to the literature: The results obtained using the proposed method and baseline approaches on the test subset are presented in Table I. Although using the vanilla KD [15] yields the best results when using $5\times$ augmentation images, its performance deteriorates as the resolution decreases and the information becomes more distorted. In addition, feature regularization via FM [4] or SR [17] terms does not show relevant improvements over KD in inter-resolution knowledge distillation. In contrast, the proposed attention matching (AT^+) methodology performs comparably to these methods at $5\times$ magnification and outperforms previous literature by $\sim 3\%$ accuracy for images at $2.5\times$, and $1.25\times$ magnification. In addition, the proposed AT^+ formulation can obtain comparable accuracy to the Teacher model, trained at $10\times$ magnification, with even 8 times less resolution.

TABLE I
COMPARISON TO PREVIOUS LITERATURE ON SICAPV2. THE METRIC PRESENTED IS THE ACCURACY AND QUADRATIC KAPPA (ACC/κ).

Method	Augmentation			
	10x	5x	2.5x	1.25x
Teacher	0.733/0.829	0.723/0.781	0.710/0.789	0.700/0.772
KD [15]	-	0.778/0.845	0.743/0.803	0.700/0.772
FM + KD [4]	-	0.741/0.828	0.733/0.806	0.709/0.787
FM + SR [17]	-	0.754/0.841	0.742/0.799	0.694/0.725
FM + AT^+	-	0.763/0.837	0.770/0.807	0.731/0.803

In the following, we provide comprehensive ablation experiments to validate the different elements of the proposed methodology. It is worth mentioning that the results further presented for the ablation experiments are obtained on the validation set.

Optimizing AT^+ distillation: First, we study the best way of leveraging normalized attention maps for knowledge distillation. To this end, after the gradient-weight of feature maps in Eq. 4 to obtain pixel-level logits, we study different settings: no normalization, min-max normalization to clip logits to $[0, 1]$ values, and min-max normalization after ReLU activation, which weights only regions with positive gradients for each target class, as used in Grad-CAM [6]. Results are depicted in Figure 2. The results in the validation show that using ReLU activation is essential for the correct knowledge distillation at lower resolutions, with improvements over other attention computations such as the one used in [18]. Also, normalizing the obtained logits improves the model optimization. This can be explained due to it allows both the Teacher and the Student to reach absolute values for the embedding space without affecting the knowledge transfer. It is worth mentioning that this term is only applied during training, so the proposed formulation does not involve any additional computational burden during inference.

Combination with other knowledge distillation terms: Next, we depict studies to evaluate the combination of the proposed attention matching (AT^+) loss with prior popular terms on knowledge distillation. Concretely, the results obtained in the validation subset of combining AT^+ with vanilla

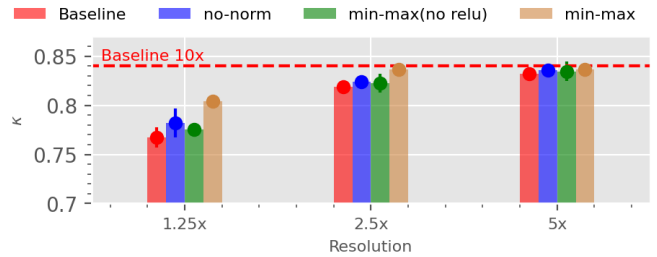


Fig. 2. Ablation study of the effect of attention map normalization on the method performance. The Teacher model trained at the different resolutions is used as a baseline.

knowledge distillation (KD) and feature matching (FM) are presented in Table II. Results show how the FM term is the only one that produces improvements over the AT^+ criteria alone, which is accentuated at lower augmentations.

TABLE II
ABLATION EXPERIMENT ON THE EFFECT OF THE COMBINATION OF DIFFERENT KNOWLEDGE DISTILLATION TERMS. THE METRIC PRESENTED IS THE ACCURACY AND QUADRATIC KAPPA (ACC/κ).

Method	Augmentation		
	5x	2.5x	1.25x
Teacher	0.746/0.831	0.740/0.818	0.706/0.767
AT^+	0.773/0.836	0.765/0.836	0.764/0.804
AT^+ + KD	0.753/0.846	0.761/0.834	0.729/0.781
AT^+ + FM	0.767/0.849	0.776/0.837	0.753/0.812
AT^+ + KD + FM	0.761/0.840	0.775/0.836	0.749/0.789

Qualitative evaluation: Finally, we include qualitative visualization of the effect of including the proposed attention matching in the training of the low-resolution Student, see Figure 3. In the context of tumor grading, for a test case with Gleason grade 3, the Teacher model’s attention (shown in the second row on the left) focuses on small individual glands. However, as the input image quality deteriorates, the model shifts its focus to different patterns, resulting in incorrect classifications. Upon incorporating the AT^+ term in the Student training (shown in the third row), the model’s attention aligns with that of the Teacher model, focusing on the same glandular patterns while simultaneously improving the output score for the correct class.

V. CONCLUSIONS

This work introduces a novel constrained formulation for knowledge distillation, aimed at enabling efficient image classification at lower resolutions. Specifically, we propose to distill the knowledge from a high-resolution Teacher model on the localization of discriminative regions in images for the given task. To achieve this, we formulate the attention matching loss, AT^+ , which forces the Student model trained at low resolution to focus on the same regions as the Teacher model by minimizing the l_2 -distance between attention maps obtained using Grad-CAMs. The proposed approach is successfully validated in the context of histology image grading. In this field, the large size of digitized biopsies and the large augmentation required are a burden for applying deep learning solutions on real-time computer-aided diagnostic systems. The obtained

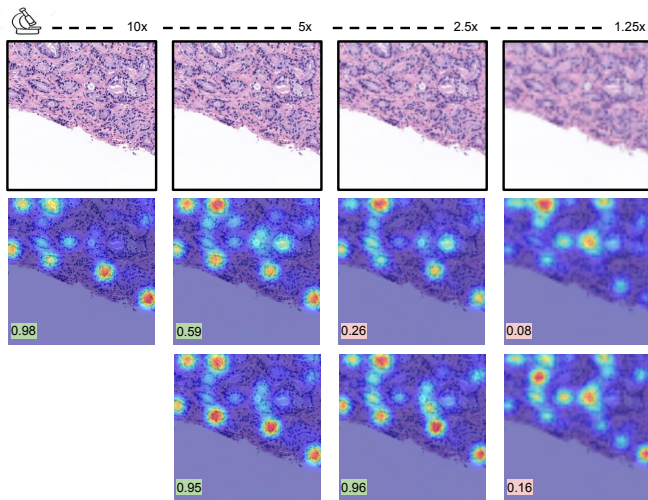


Fig. 3. Qualitative assessment of the effect of the attention matching (AT⁺) term. The top row presents original images at different resolution levels (augmentations). The second row shows the Student output to the target class trained without any knowledge distillation and the attention map produced. The last row shows the effect of distilling the knowledge from the Teacher model trained at 10x magnification. Green probabilities indicate a correctly classified sample, while red indicates the opposite.

results show that attention distillation allows operating at up to 8× fewer augmentations and outperforms previous relevant literature in knowledge distillation in ~3% accuracy. Still, it is worth mentioning that the performance of these methods might be limited at some minimum resolution since any of the baselines and proposed methods performed properly at resolutions below 0.625× augmentations. We believe that the results obtained open further research directions to allow the efficient use of deep learning solutions in the medical context.

ACKNOWLEDGEMENT

This work has received funding from Horizon 2020, the European Union’s Framework Programme for Research and Innovation, under the grant agreement No. 860627 (CLARIFY), the Spanish Ministry of Economy and Competitiveness through project PID2019-105142RB-C21 (AI4SKIN) and GVA through the project INNEST/ 2021/321 (SAMUEL). The work of Rocío del Amor has been supported by the Spanish Ministry of Universities (FPU20/05263). The work of J. Silva-Rodríguez was carried out during his previous position at Universitat Politècnica de València. The work of Adrián Colomer has been supported by the ValgrAI – Valencian Graduate School and Research Network for Artificial Intelligence and Generalitat Valenciana and Universitat Politècnica de Valencia (PAID-PD-22).

REFERENCES

[1] G. Campanella, M. G. Hanna, L. Geneslaw, A. Mirafior, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images,” *Nature Medicine*, vol. 25, pp. 1301–1309, 8 2019.

[2] J. Silva-Rodríguez, A. Colomer, and V. Naranjo, “Weglenet: A weakly-supervised convolutional neural network for the semantic segmentation of gleason grades in prostate histology images,” *Computerized Medical Imaging and Graphics*, vol. 88, pp. 1–10, 3 2021.

[3] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, “Mitosis detection in breast cancer histology images with deep neural networks,” *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 411–418, 2013.

[4] J. DiPalma, A. A. Suriawinata, L. J. Tafe, L. Torresani, and S. Hasanpour, “Resolution-based distillation for efficient histology image classification,” *Artificial Intelligence in Medicine*, vol. 119, 9 2021.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, vol. 128, 10 2017.

[7] D. Pathak, P. Krähenbühl, and T. Darrell, “Constrained convolutional neural networks for weakly supervised segmentation,” *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–12, 6 2015.

[8] J. Silva-Rodríguez, A. Schmidt, M. A. Sales, R. Molina, and V. Naranjo, “Proportion constrained weakly supervised histopathology image classification,” *Computers in Biology and Medicine*, vol. 147, 2022.

[9] R. del Amor, P. Meseguer, T. L. Parigi, V. Villanacci, A. Colomer, L. Launet, A. Bazarova, G. E. Tontini, R. Bisschops, G. de Hertogh, J. G. Ferraz, M. Götz, X. Gui, B. H. Hayee, M. Lazarev, R. Panacione, A. Parra-Blanco, P. Bhandari, L. Pastorelli, T. Rath, E. S. Røyset, M. Vieth, D. Zardo, E. Grisan, S. Ghosh, M. Iacucci, and V. Naranjo, “Constrained multiple instance learning for ulcerative colitis prediction using histological images,” *Computer Methods and Programs in Biomedicine*, vol. 224, 9 2022.

[10] J. Silva-Rodríguez and J. Dolz, “Looking at the whole picture: constrained unsupervised anomaly segmentation,” *British Machine Vision Conference (BMVC)*, 2021.

[11] J. Silva-Rodríguez, V. Naranjo, and J. Dolz, “Constrained unsupervised anomaly segmentation,” *Medical Image Analysis*, vol. 80, 2022.

[12] S. Venkataramanan, K. C. Peng, R. V. Singh, and A. Mahalanobis, “Attention guided anomaly localization in images,” *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[13] R. R. Selvaraju, K. Desai, J. Johnson, and N. Naik, “Casting your model: Learning to localize improves self-supervised representations,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[14] M. Sun, Y. Yuan, F. Zhou, and E. Ding, “Multi-attention multi-class constraint for fine-grained image recognition,” *Proceedings of the European Conference on Computer Vision (ECCV)*, 6 2018.

[15] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *Advances in Neural Information Processing Systems (NeurIPS)*, 3 2015.

[16] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “Fitnets: Hints for thin deep nets,” *International Conference on Learning Representations (ICLR)*, 12 2015.

[17] J. Yang, B. Marinéz, A. Bulat, and G. Tzimiropoulos, “Knowledge distillation via softmax regression representation learning,” *International Conference on Learning Representations (ICLR)*, 2021.

[18] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” *International Conference on Learning Representations (ICLR)*, 12 2017.

[19] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, “Self-training with noisy student improves imagenet classification,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11 2020.

[20] M. Hu, M. Maillard, Y. Zhang, T. Ciceri, G. Barbera, I. Bloch, P. Gori, and G. L. Barbera, “Knowledge distillation from multi-modal to mono-modal segmentation networks,” *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020.

[21] J. Silva-Rodríguez, A. Colomer, M. Sales, R. Molina, and V. Naranjo, “Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection,” *Computer Methods and Programs in Biomedicine*, vol. 195, 2020.