Background-Invariant Independence-Guided Multi-head Attention Network for Skin Lesion Classification

Debasmit Roy[1⋆], Srinjoy Dutta[1], Soham Bose[1], Friedhelm Schwenker[2], and Ram Sarkar[1]

[1] Department of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal, India
{debasmitr.cse.ug, srinjoyd.cse.ug, sohamb.cse.ug, ram.sarkar}@jadavpuruniversity.in

[2] Institute of Neural Information Processing, Ulm University, James Franck Ring, 89081 Ulm, Germany
friedhelm.schwenker@uni-ulm.de

Abstract. Biomedical image classification faces several adversarial challenges, including occlusions from artifacts, variations in tissue pigmentation, and class imbalance, which hinder model generalization. Existing attention mechanisms enhance region localization but often introduce redundant dependencies across attention heads, limiting feature diversity. We propose the Background-Invariant Independence-Guided Multi-head Attention Network (BIIGMA-Net) to address these issues. BIIGMA-Net employs Multi-head Independence-Guided Channel Attention (MICA), where each head independently learns feature importance while enforcing neuron-wise independence using the Hilbert-Schmidt Independence Criterion (HSIC) to enhance feature diversity. Additionally, a saliency-driven mechanism suppresses background activations by selectively shuffling non-salient vectors, preventing the model from relying on static background cues. By integrating these strategies, BIIGMA-Net improves robustness against spurious background noise while ensuring complementary feature extraction. Extensive experiments on popular skin cancer datasets (ISIC-17, ISIC-18 and ISIC-19) demonstrate the framework's effectiveness and robustness. Our code is available at: https://github.com/shb2908/BIIGMA-Net

Keywords: Channel and spatial attention · Medical imaging · Noise regularization · Skin cancer · Skin lesion classification.

1 Introduction

Skin cancer, comprising various malignancies, is a significant global health concern. The World Health Organization (WHO) reported over 1.5 million new cases and approximately 350,000 melanoma diagnoses annually, with 57,000 deaths in

⋆ Corresponding author: debasmitr.cse.ug@jadavpuruniversity.in

2020. Early diagnosis is critical, as delayed detection increases health risks and burdens healthcare systems. Current diagnosis relies on visual inspection by dermatologists, which is subjective and time-consuming.

CNN-based automated classification methods have shown promise but face challenges in biomedical imaging. One major limitation is the model's discriminative ability, which is often compromised by adversarial factors such as hair artifacts, skin pigmentation, and variations in complexion. These introduce noise, making it difficult for models to distinguish pathological patterns from background textures. Moreover, lesions often occupy a small portion of the image, leading to ineffective feature extraction. Class imbalance further weakens the model's generalization by biasing it toward majority classes, prompting memorization rather than learning discriminative features. This reduces sensitivity to rare conditions, which are often of greater clinical significance.

Researchers have tackled two key challenges in skin lesion classification: class imbalance in multi-class datasets and diverse imaging artifacts. To address class imbalance, techniques such as up-sampling and under-sampling [10,15] have been explored. Apart from these, recent methods leverage contrastive learning and self-distillation to enhance feature representation and model generalization. Xu et al. [21] proposed a model-agnostic self-supervised knowledge distillation approach using noisy teacher predictions. Zhang et al. [25] introduced Class-Enhancement Contrastive Learning (ECL) with a hybrid-proxy model and balanced-weighted loss. Li et al. [14] developed Targeted Supervised Contrastive Learning (TSC) to enforce uniform class feature distribution for better separability. Yao et al. [23] applied deep learning with DropOut, DropBlock, RandAugment, and a multi-weighted loss to improve feature extraction. Chu et al. [5] proposed a deep-learning model with class-agnostic activation maps to enhance melanoma diagnosis under varying imaging conditions.

Popular channel attention mechanisms such as SE-Net and CBAM [12,20] introduce lightweight modules that recalibrate channel-wise feature responses by modeling interdependencies between channels. Zhang et al. [24] proposed attention residual learning blocks, using higher-layer feature maps as attention masks for lower layers to enhance feature representation. Ding et al. [7] further improved attention-based learning by leveraging class activation maps across multiple layers via matrix multiplication and concatenation. Wei et al. [19] extended this with a dual attention module, where spatial attention captures local patterns, and channel attention strengthens global feature dependencies.

Multiple-Exit CAM [4] captures activation maps at different resolutions to improve spatial attention robustness. Transformer-based attention mechanisms, such as TransAttUnet [2], have also been integrated into image classification and segmentation models. However, whether attention heads learn complementary information remains an open question, as explored in [1,3]. Additionally, adversarial robustness has been studied through various approaches [22,16], addressing artifacts like hair and pigmentation in skin lesion and biomedical images.

In our proposed Background-Invariant Independence-Guided Multi-head Attention (BIIGMA-Net), we enforce independence across projection heads in the

channel attention block by minimizing mutual information, reducing redundancy in feature representation. To improve robustness against spurious background noise, we generate a hybrid feature map by selectively shuffling non-salient vectors using an inverted saliency map, ensuring classification consistency. To the best of our knowledge, this is the first work integrating an independence criterion in CNN-based attention heads alongside background agnosticism.The key contributions of our work are:

1. **Multi-head Independence-Guided Channel Attention (MICA):** We introduce a multi-head channel attention mechanism where each head independently learns feature importance. To enforce decorrelation, we use the Hilbert-Schmidt Independence Criterion (HSIC) at the neuron level instead of covariance matrices, which fail to capture higher-order dependencies. This ensures diverse and complementary feature extraction, reducing redundancy and more information count in the final representation.
2. **Spatial Attention Guided Background Invariance:** To suppress irrelevant background features while preserving discriminative information, we employ a saliency-driven mechanism that samples and shuffles background feature vectors. This prevents reliance on static background cues, enhancing robustness against background variations and spurious correlations.

2 Proposed Method

In our proposed method, we focus on two aspects. Firstly, Channel Attention Block is introduced to enhance non-redundant information by reducing mutual information across feature projections using our independence criterion. Secondly, Spatial Attention Guided Vector Sampling aggregates shallow convolutional features into a saliency map and filters non-salient regions, enforcing background-invariant learning through adversarial background shuffling.

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{m}$, the final prediction is $\hat{y} = F \circ G(x_i)$, where F is the convolutional backbone and G is the classification head. The final convolutional feature block is $R_i^{\text{last}} = F(x_i)$, where $R_i^{\text{last}} \in \mathbb{R}^{H \times W \times C}$. Additionally, convolutional features from shallower depths are denoted as $\{R_i^{(l)}\}_{l=1}^{\#\text{early exits}}$. R^{last} is represented as R in later section for simplicity.

2.1 Multi-head Independence Guided Channel Attention

Our goal is to extract mutually independent projections from the original convolutional feature set, increasing non-redundant information in R^{last}. We achieve this by passing the features through K parallel projection blocks (P, 1×1 Conv + SE Block), generating K projected feature sets $\{R_{proj}^k\}_{k=1}^{K}$. The feature block is then reshaped as follows:

$$\mathbb{R}^{B \times H \times W \times C} \rightarrow \mathbb{R}^{B \times HW \times C}$$

yielding $\{r^k\}_{k=1}^{K}$, which can be viewed as K sets of HW C-dimensional vectors for a single sample in the batch.

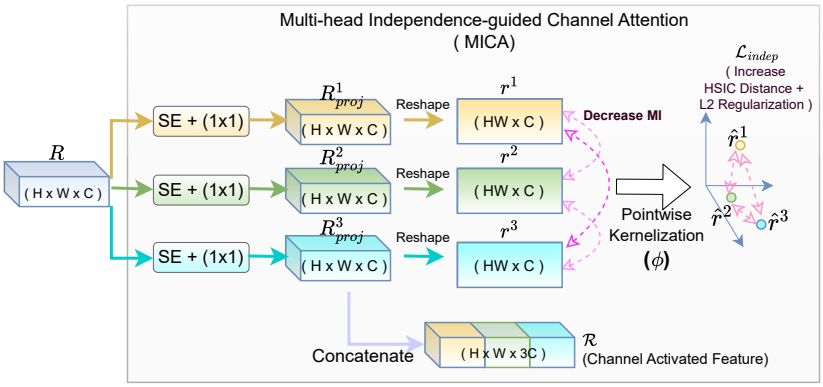

Fig. 1. Proposed MICA block projects the input feature into multiple subspaces. HSIC criterion is minimized to reduce mutual information among these projections.

We impose independence between neuron activations across different heads. Specifically, we minimize the Hilbert-Schmidt Independence Criterion (HSIC) [9,8] for each pair $\{r^{k_1}, r^{k_2}\}_{k_1 \neq k_2}$ within each batch. HSIC first maps the covariance matrix of mean-normalized features $\{\hat{r}^{k_1}, \hat{r}^{k_2}\}$ into kernel space using an RBF kernel for each HW number of spatial locations:

$$\mathcal{C}_{\{\hat{r}^{k_1},\hat{r}^{k_2}\}} = \frac{1}{HW}\sum_{j=1}^{HW}(\hat{r}_j^{k_1}\cdot\hat{r}_j^{k_2}) = \frac{1}{HW}(\hat{r}_1^{k_1^T}\cdot C\cdot\hat{r}^{k_2}) \xrightarrow[rbf]{} \frac{1}{HW}(\phi_{r^{k_1}}^T\cdot C\cdot\phi_{r^{k_2}}) \tag{1}$$

$$\begin{aligned} HSIC(\{r^{k_1}, r^{k_2}\}) = ||\mathcal{C}_{\{\hat{r}^{k_1},\hat{r}^{k_2}\}}||_F &= \frac{1}{(HW)^2}Tr[\phi_{r_{k_1}}\phi_{r_{k_1}}^T C\phi_{r_{k_2}}\phi_{r_{k_2}}^T C] \\ &= \frac{1}{(HW)^2}Tr[K_{r_{k_1}} C K_{r_{k_2}} C] \end{aligned} \tag{2}$$

Here C is centering matrix and ϕ is the kernelized feature. Kernelization accounts for dependencies across different neuron indices, whereas covariance considers only identical indices. We optimize the independence criterion across all head pairs while applying L_2 regularization ($\lambda = 1$) on projection head weights to mitigate overfitting:

$$\mathcal{L}_{indep} = \frac{1}{K^2}\sum_{k_i}^{K}\sum_{k_i \neq k_j}^{K} HSIC(\hat{r}^{k_i}, \hat{r}^{k_j}) + \lambda l_2^{norm}(W_{heads}) \tag{3}$$

Finally, we concatenate features from all heads to obtain the final channel activated (CA) features:

$$\mathcal{R}^{\text{CA}} = [R_{proj}^k]_{k=1}^K \tag{4}$$

where [.] denotes channel-wise concatenation.

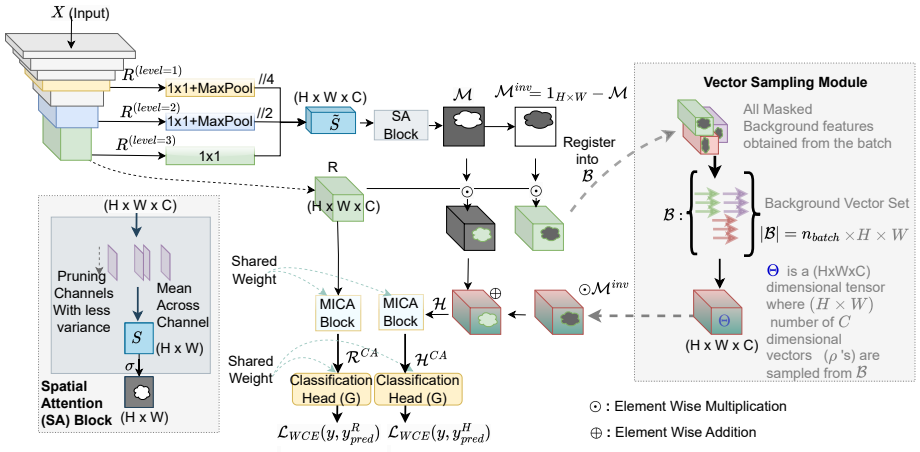

Fig. 2. Proposed BIIGMA Network. The bottom-left shows the spatial attention block, which processes a feature block to generate an attention map. The right side depicts the vector sampling mechanism, storing batch-wise feature blocks and selectively sampling the background vector to form a hybrid feature.

2.2 Spatial Attention Guided Vector Sampling

We extract convolutional features from early exits (shallow depths of the backbone) and pass each through adapters (A) comprising a 1×1 layer and a max-pooling layer. The processed features are concatenated as $\tilde{S} = [A(R_i^{(l)})]_{l=1}^{\#\text{early exits}}$, where [.] denotes concatenation. Max-pooling ensures consistent spatial dimensions. Channel pruning is then applied by removing low-variance channels [26,13] to mitigate overestimation of point-wise spatial attention scores, yielding the pruned feature:

$$S = [\tilde{S}_j]_{j=1, var(\tilde{S}_j) < mean(var(\tilde{S}))}^{\#channels}$$

The spatial attention (SA) map is computed as the channel-wise mean followed by sigmoid activation:

$$\mathcal{M} = SA(S) = \text{sigmoid}(\text{mean}(S, \text{channel axis})), \quad \mathcal{M} \in \mathbb{R}^{H \times W} \tag{5}$$

We define,

$$\mathcal{M}^{inv} = \mathbf{1}_{H \times W} - \mathcal{M} \tag{6}$$

Here, $\mathcal{M}^{inv}$ represents the non-salient region. The background set $\mathcal{B}$ is filtered from the last convolutional feature R:

$$\{R_i \odot \mathcal{M}_i^{inv}\}_{i=1}^{n_{batch}} = \{\overrightarrow{\rho}_j\}_{j=1 \text{ and } \overrightarrow{\rho}_j \neq \overrightarrow{0}}^{n_{batch} \times H \times W}$$

where $\overrightarrow{\rho}$ represents point-wise $\mathbb{R}^C$-dimensional vectors. A sampled background tensor Θ of shape $\mathbb{R}^{H \times W \times C}$ is generated by randomly sampling the

vectors from $\mathcal{B}$ for each example. The hybrid feature map for i^{th} sample is then constructed as:

$$\mathcal{H}_i = \mathcal{M}_i \odot R_i + \mathcal{M}_i^{inv} \odot \Theta_i$$

Finally, $\mathcal{H}_i$ is passed through the proposed the same MICA block:

$$\mathcal{H}^{CA} = \text{MICA}(\mathcal{H}) \tag{7}$$

This sampling enhances MICA's robustness to background cues by enforcing invariance criteria.

2.3 Overall Loss Function

We define the classification objective as

$$\mathcal{L}_{classif} = \mathcal{L}_{WCE}(G(\mathcal{R}^{CA}), Y) + \mathcal{L}_{WCE}(G(\mathcal{H}^{CA}), Y) \tag{8}$$

where $\mathcal{L}_{WCE}$ denotes the Weighted Cross-Entropy Loss where class weights are estimated as the inverse class frequency. Minimizing the $\mathcal{L}_{WCE}$ for original branch prediction i.e. $\mathcal{R}^{CA}$ as well as the sampling branch prediction $\mathcal{H}^{CA}$ ensures the invariance to background cues. The final loss is

$$\mathcal{L} = \mathcal{L}_{classif} + \mathcal{L}_{indep} \tag{9}$$

During inference, sampling is skipped, and predictions are taken from the original branch only.

3 Experiments and Results

3.1 Datasets and Experimental Setup

We utilize three widely recognized dermoscopic datasets: ISIC-17 [6], ISIC-18 [17], and ISIC-19 [11]. ISIC-17 comprises 2,600 images across three classes: Nevus (NV), Seborrheic Keratosis (SK), and Melanoma (MEL). ISIC-18 contains 10,015 images spanning seven classes: MEL, NV, Basal Cell Carcinoma (BCC), Actinic Keratosis (AKIEC), Benign Keratosis (BKL), Dermatofibroma (DF), and Vascular Lesion (VASC). ISIC-19 includes 25,331 images with an additional class, Squamous Cell Carcinoma (SCC). All datasets exhibit severe class imbalance (imbalance factor > 50) and are split into an 80:20 train-test ratio. For ISIC-17, we perform binary classification for Melanoma vs. others and Seborrheic Keratosis vs. others, along with multiclass classification.

Each image is resized to (224×224) and normalized by $\frac{1}{255}$. We employ an ImageNet-pretrained DenseNet-121 backbone, trained using the Adam optimizer (learning rate: 0.0001) for up to 100 epochs. Spatial attention-guided vector sampling is activated after 20 epochs to mitigate early adversarial effects. Experiments are conducted on Nvidia Tesla T4 GPUs (28GB RAM) using Python 3.9 and TensorFlow Keras. Performance is evaluated using accuracy, macro-averaged F1-score, recall, and precision.

Table 1. Performance comparison across different ISIC datasets.

Dataset	Method	F1	Acc	Rec	Prec
ISIC 17	Zhang et al. [24] (Mel vs Others)	-	0.8370	0.5900	-
	Zhang et al. [24] (SK vs Others)	-	0.9080	0.7780	-
	Ding et al. [7] (Mel vs Others)	-	0.8620	0.6320	-
	Ding et al. [7] (SK vs Others)	-	**0.9280**	0.8330	-
	Wei et al. [19] (Mel vs Others)	-	**0.8620**	0.6620	-
	Ours (Mel vs Others)	**0.7933**	0.8486	**0.7675**	**0.7001**
	Ours (SK vs Others)	**0.8044**	0.9092	**0.8428**	**0.7784**
ISIC 18	Li et al. [14]	0.7494	0.8594	0.7335	0.7777
	Zhang et al. [25]	0.7676	0.8720	0.7301	0.8344
	Wang et al. [18]	0.7952	0.8542	0.8557	0.7320
	Xu et al. [21]	0.7968	0.8442	**0.8595**	0.7499
	Ours	**0.8220**	**0.8833**	0.8094	**0.8351**
ISIC 19	Li et al. [14]	0.7513	0.8475	0.7189	0.7981
	Yao et al. [23]	0.7508	0.8410	0.7483	0.7581
	Chu et al. [5]	0.7920	0.8820	**0.8700**	-
	Zhang et al. [25]	0.7946	0.8611	0.7657	**0.8322**
	Ours	**0.8052**	**0.8853**	0.8130	0.7975

Table 2. Ablation study on ISIC datasets with different settings.

Setting	ISIC 17			ISIC 18			ISIC 19		
	F1	Acc	Rec	F1	Acc	Rec	F1	Acc	Rec
Baseline	0.6408	0.7187	0.6311	0.7256	0.8431	0.7221	0.7178	0.8022	0.7334
W/O MICA	0.6957	0.7684	0.6947	0.7994	0.8798	0.7931	0.7945	0.8722	0.7802
W/O Vector Sampling	0.6994	0.7633	0.7984	0.7947	0.8638	0.8028	0.7988	0.8760	0.8051
W/O $\mathcal{L}_{\text{indep}}$	0.6860	0.7450	0.7094	0.7869	0.8738	0.7627	0.7900	0.8685	0.7609
Heads = 1	0.6896	0.7600	0.6956	0.7871	0.8691	0.7789	0.7892	0.8701	0.7758
Heads = 2	0.6943	0.7733	0.6872	0.7938	0.8753	0.7865	0.7952	0.8775	0.7803
Heads = 3	0.7203	0.7921	0.7116	0.8220	0.8833	0.8094	0.8025	0.8814	0.8086
Heads = 4	0.7153	0.7867	0.7160	0.7920	0.8743	0.7781	0.7961	0.8780	0.7829
Sigma = 0.5	0.6700	0.7400	0.6908	0.7883	0.8748	0.7634	0.7832	0.8690	0.7605
Sigma = 1.0	0.6858	0.7567	0.6979	0.8220	0.8833	0.8094	0.8052	0.8853	0.8130
Sigma = 2.0	0.7203	0.7921	0.7116	0.7937	0.8783	0.7842	0.7904	0.8731	0.7760
Sigma = 4.0	0.6579	0.7381	0.6606	0.7866	0.8748	0.7731	0.7995	0.8802	0.7922
Cosine Similarity based $\mathcal{L}_{\text{indep}}$	0.7133	0.7883	0.7063	0.7920	0.8753	0.7817	0.7941	0.8726	0.7779
Best Configuration	**0.7203**	**0.7921**	**0.7116**	**0.8220**	**0.8833**	**0.8094**	**0.8052**	**0.8853**	**0.8130**

3.2 Analysis of Results

In Table 1, the proposed framework outperforms existing methods, with an average increase of 1-2% in F1-score, which is the most reliable metric in highly imbalanced settings.

Table 2 shows ablation study results with different configurations. First, we evaluate the performance of the backbone without enhancements, where the baseline achieves an F1-score of 0.7256 on ISIC 18. Gradually increasing the number of heads from 1 to 4 results in a performance improvement. The highest performance is observed with 3 heads, yielding an F1-score improvement of 9.6% over the baseline. More heads enrich feature representation but introduce two issues: increased feature complexity in the post-CNN phase, leading to overfitting, and redundant features across heads. To address this, we experiment with and without the $\mathcal{L}_{\text{indep}}$ loss, observing a 3.3% improvement in F1-score on ISIC 18.

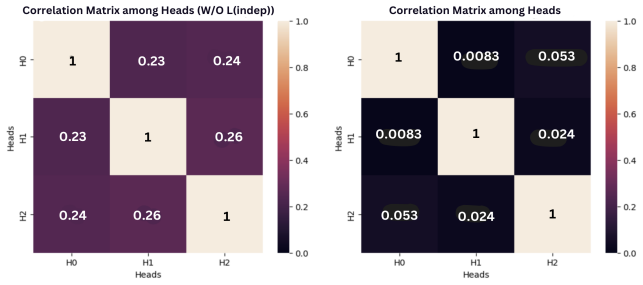

Fig. 3. Cross-correlation matrix representing the mean cosine similarity among head-level features (trained on ISIC-18). The left matrix corresponds to training without the independence criterion, while the right one includes it.

We qualitatively analyze the pairwise cross-correlation among all attention heads. In Figure 3, we observe a strong correlation between heads when trained without $\mathcal{L}_{indep}$. In contrast, when the head-level features are decorrelated, each head captures distinct information. To clarify the essence of the HSIC criterion, rather than using a naive dot product-based correlation, we replace HSIC with cosine similarity and observe a steep decrease in performance (0.8220 to 0.7920 in terms of F1). HSIC captures higher-order correlations, accounting for the covariance of similar neurons even if they reside at different neuron indices. We use the RBF kernel to transform features, where the bandwidth σ is critical. A smaller σ makes the kernel sensitive to minor variations, while a larger σ smooths the dependency structure. We conduct extensive experiments with different σ values to select the optimal one for each dataset ($\sigma_{opt} = 2.0$, $\sigma_{opt} = 1.0$ & $\sigma_{opt} = 1.0$ are optimal choices for ISIC-17,18 & 19 datasets respectively), as it depends on the feature-level variance. Furthermore, to demonstrate the contribution of spatial attention guidance, we observe a clear increase in F1 scores when combined with feature-level augmentation (2.5%, 1.5%, 1% respectively in each dataset).

4 Conclusion

We propose BIIGMA-Net, integrating MICA module to enforce independence across projection heads by minimizing mutual information, reducing feature redundancy. To mitigate spurious background noise in dermoscopic datasets, we generate a hybrid feature map by selectively shuffling non-salient vectors using an inverted saliency map, ensuring classification consistency. Novelty of our work relies on combining an independence criterion in CNN-based attention heads with background agnosticism. Experiments on ISIC-17, ISIC-18, and ISIC-19 confirm its effectiveness. In the future, we can include adaptive frameworks for kernel bandwidth that dynamically adjust to feature distributions and investigating methods to enhance robustness against adversarial perturbations.

Acknowledgments. This research work was supported by the departmental project titled "FIST Engineering B/C/D Project", funded by the Department of Science

and Technology (DST), Government of India, under reference number SR/FST/ET-I/2022/1059(C).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bian, Y., Huang, J., Cai, X., Yuan, J., Church, K.: On attention redundancy: A comprehensive study. In: Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies. pp. 930–945 (2021)
2. Chen, B., Liu, Y., Zhang, Z., Lu, G., Kong, A.W.K.: Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. IEEE Transactions on Emerging Topics in Computational Intelligence **8**(1), 55–68 (2023)
3. Chen, T., Zhang, Z., Cheng, Y., Awadallah, A., Wang, Z.: The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12020–12030 (2022)
4. Chen, Y.J., Hu, X., Shi, Y., Ho, T.Y.: Ame-cam: Attentive multiple-exit cam for weakly supervised segmentation on mri brain tumor. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 173–182. Springer (2023)
5. Chu, Y., Lee, S., Oh, B., Yang, S.: Class-agnostic feature-learning-based deep-learning model for robust melanoma prediction. IEEE Journal of Biomedical and Health Informatics (2025)
6. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
7. Ding, S., Wu, Z., Zheng, Y., Liu, Z., Yang, X., Yang, X., Yuan, G., Xie, J.: Deep attention branch networks for skin lesion classification. Computer methods and programs in biomedicine **212**, 106447 (2021)
8. Greenfeld, D., Shalit, U.: Robust learning with the Hilbert-schmidt independence criterion. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 3759–3768. PMLR (13–18 Jul 2020), `https://proceedings.mlr.press/v119/greenfeld20a.html`
9. Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., Smola, A.: A kernel statistical test of independence. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems. vol. 20. Curran Associates, Inc. (2007)
10. HaCohen, Y., Fattal, R., Lischinski, D.: Image upsampling via texture hallucination. In: 2010 IEEE International Conference on Computational Photography (ICCP). pp. 1–8. IEEE (2010)
11. Hernández-Pérez, C., Combalia, M., Podlipnik, S., Codella, N.C., Rotemberg, V., Halpern, A.C., Reiter, O., Carrera, C., Barreiro, A., Helba, B., et al.: Bcn20000: Dermoscopic lesions in the wild. Scientific data **11**(1), 641 (2024)

12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745
13. Khetan, A., Karnin, Z.: Prunenet: Channel pruning via global importance. arXiv preprint arXiv:2005.11282 (2020)
14. Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R.S., Indyk, P., Katabi, D.: Targeted supervised contrastive learning for long-tailed recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6918–6928 (2022)
15. Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P.: Fully convolutional neural networks for remote sensing image classification. In: 2016 IEEE international geoscience and remote sensing symposium (IGARSS). pp. 5071–5074. IEEE (2016)
16. Nguyen-Duc, T., Le, T., Bammer, R., Zhao, H., Cai, J., Phung, D.: Cross-adversarial local distribution regularization for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 183–194. Springer (2023)
17. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
18. Wang, Y., Wang, Y., Cai, J., Lee, T.K., Miao, C., Wang, Z.J.: Ssd-kd: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. Medical Image Analysis **84**, 102693 (2023)
19. Wei, Z., Li, Q., Song, H.: Dual attention based network for skin lesion classification with auxiliary learning. Biomedical Signal Processing and Control **74**, 103549 (2022)
20. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 3–19. Springer International Publishing, Cham (2018)
21. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: European conference on computer vision. pp. 588–604. Springer (2020)
22. Xu, Y., Xie, S., Reynolds, M., Ragoza, M., Gong, M., Batmanghelich, K.: Adversarial consistency for single domain generalization in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 671–681. Springer (2022)
23. Yao, P., Shen, S., Xu, M., Liu, P., Zhang, F., Xing, J., Shao, P., Kaffenberger, B., Xu, R.X.: Single model deep learning on imbalanced small datasets for skin lesion classification. IEEE transactions on medical imaging **41**(5), 1242–1254 (2021)
24. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Attention residual learning for skin lesion classification. IEEE transactions on medical imaging **38**(9), 2092–2103 (2019)
25. Zhang, Y., Chen, J., Wang, K., Xie, F.: Ecl: Class-enhancement contrastive learning for long-tailed skin lesion classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 244–254. Springer (2023)
26. Zhao, C., Ni, B., Zhang, J., Zhao, Q., Zhang, W., Tian, Q.: Variational convolutional neural network pruning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2780–2789 (2019)