MEDIATOR: Enhancing Medical Diagnosis via Gated Distillation and Decoupled Learning

Kyuan Oh*, Sunguk Lee*, Mina Hwang*, Heewon Yang*, Kiseong Lee†
Department of Artificial Intelligence, AI Humanities Research Institute†*
Chung-Ang University
Seoul, Republic of Korea
oka04108@cau.ac.kr, bill8342@cau.ac.kr, mh0506@cau.ac.kr, heewon6205@cau.ac.kr, goory@cau.ac.kr

***Abstract*—The integration of multimodal data, such as medical images and structured clinical records, is crucial for enhancing diagnostic accuracy. However, effectively fusing these heterogeneous data types poses significant challenges, primarily the semantic gap between modalities and the risk of feature degradation in joint training. To address these issues, we propose MEDIATOR (Mutual-information Enhanced DIstillation and gAting for Two-phase Orchestrated Representation), a novel learning framework that fundamentally decouples the learning process into two distinct phases: Phase 1 focuses exclusively on learning a powerful, unified vision representation from heterogeneous imaging data. It employs a Mutual information Gating and Distillation (MGD) mechanism to effectively bridge the modality gap and generate fusion-optimized features. In Phase 2, the entire vision module remains frozen, and a lightweight classifier is trained to integrate these high-quality visual features with auxiliary tabular data. This decoupled strategy boosts diagnostic performance by augmenting the rich visual representations with supplementary clinical information from tabular data, providing a holistic view of the diagnosis while preserving the integrity of visual features. Our experiments show that this paradigm significantly enhances diagnostic performance, establishing a robust and effective approach for multimodal medical diagnosis.**

***Index Terms*—Multimodal Representation Learning, Medical Diagnosis, Decoupled Learning, Feature Fusion, Knowledge Distillation, Information Gating**

I. Introduction

Multimodal deep learning, which integrates heterogeneous sources such as radiological images and structured patient records, holds immense promise for advancing medical diagnosis [1]. By leveraging complementary views, models can potentially achieve a more holistic understanding, leading to more accurate and robust clinical predictions. However, the path to effective fusion is fraught with challenges.

A central obstacle is the *modality gap*—the large discrepancy in statistical properties and dimensionality between different data types, which complicates joint representation learning [2]. A further pitfall of end-to-end fusion is *representational dilution*: when complex imaging data and simpler tabular data are trained together, the learning process can be dominated by the modality with stronger or more easily learnable signals. This not only washes out rich visual patterns but can also lead to spurious correlations, undermining generalization [3], [4].

We address these challenges by departing from the conventional end-to-end paradigm. Instead, we propose a **decoupled, two-phase learning framework named MEDIATOR**. In Phase 1, the model is dedicated to building a strong visual foundation through a Mutual-Information Gated Distillation (MGD) mechanism that balances contributions across imaging modalities. In Phase 2, the vision module remains frozen, and auxiliary tabular data are integrated via a lightweight classifier, allowing complementary signals to enhance rather than interfere.

Our contributions are threefold:

- **Framework:** We introduce MEDIATOR, a decoupled two-phase paradigm that separates high-capacity visual representation learning from auxiliary integration.
- **Mechanism:** We design Mutual information Gating and Distillation (MGD), a gating–distillation strategy that produces fusion-optimized visual features while mitigating modality dominance.
- **Empirical Validation:** We demonstrate that our decoupled approach with a frozen vision module significantly outperforms standard end-to-end fusion techniques, achieving high performance on a challenging multimodal medical diagnosis task.

II. Related Work

A. Multimodal Learning in Medicine

Clinical practice is inherently multimodal, relying on images, laboratory tests, and patient records. Yet, most medical AI systems remain unimodal, leaving multimodal integration underexplored. Recent studies have shown that combining imaging with structured clinical data improves diagnosis, prognosis, and risk stratification [1], [4], [5]. These results highlight the pressing need for robust fusion frameworks that can handle heterogeneity and alignment challenges in real-world clinical data [6].

B. Evolution of Fusion Architectures

Early fusion methods, such as concatenating features or averaging predictions [7] provided initial benefits but lacked the capacity to model complex cross-modal interactions. This motivated intermediate or joint fusion architectures, which learn shared representation spaces for deeper modality interaction [3]. While effective, these designs remain constrained

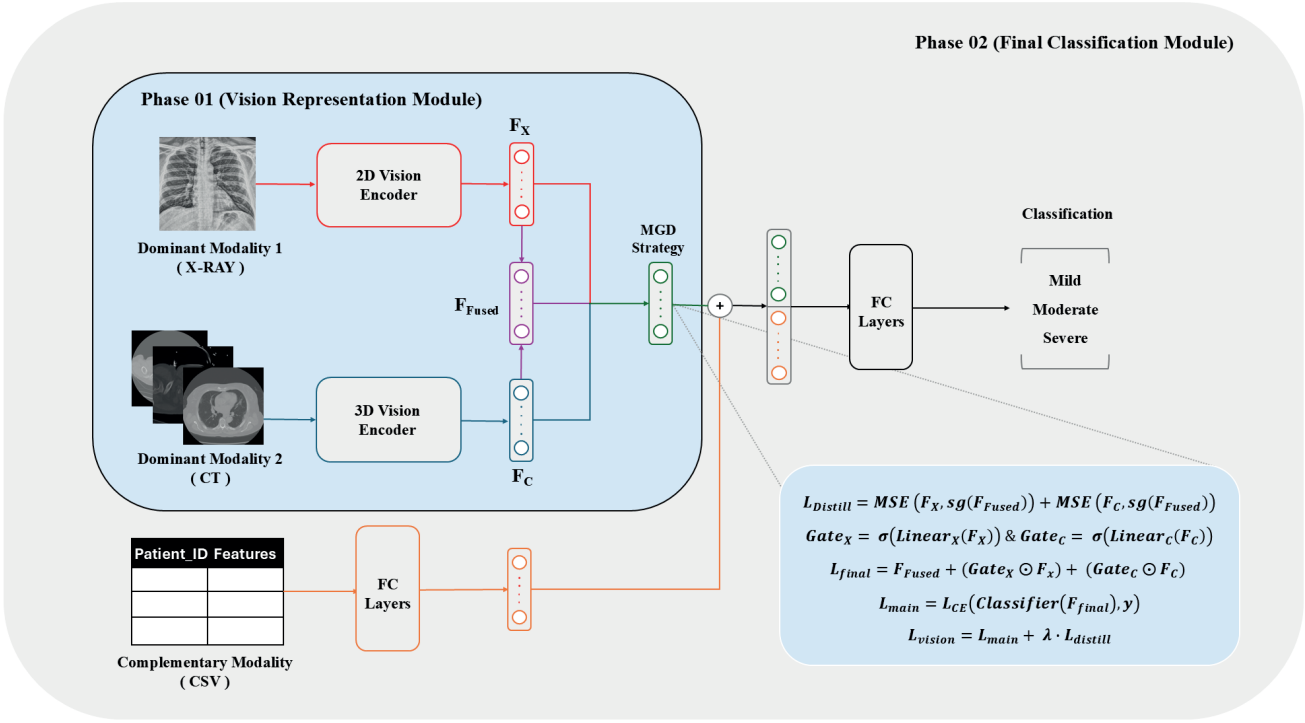

Fig. 1. **Overview of the proposed two-phase framework, MEDIATOR.** Phase 1 (Vision Representation): Two imaging modalities (X-ray, CT) are encoded by separate backbones and fused via the proposed MGD strategy to obtain a robust shared representation. Phase 2 (Final Classification): The vision module remains frozen. Features are integrated with auxiliary tabular data processed by a lightweight MLP, followed by a fusion block and classifier to produce the final diagnosis.

by end-to-end training, where the optimization process can blur modality-specific information. Our work departs from this trajectory by introducing a decoupled training protocol that avoids such representational dilution.

C. Strategies for Bridging the Modality Gap

A central challenge in multimodal learning is the *modality gap*—the mismatch between modalities in scale, structure, and semantics. Several strategies have been explored:

1) Embedding Alignment: Contrastive and distributional approaches project modalities into a common latent space [8]. Large-scale vision–language models such as CLIP [9], and their medical adaptations ConVIRT and MedCLIP [10], [11], demonstrate the power of alignment.

2) Cross-Modal Attention: Transformer-based attention enables modalities to selectively query each other [12], and has been applied to medical tasks [13]. While flexible, attention can be computationally heavy and prone to bias toward dominant modalities.

3) Knowledge Distillation: Distillation transfers knowledge across modalities or from fused "teacher" models to individual "student" encoders [14], [15], regularizing them to produce fusion-friendly features.

D. Research Gap and Our Approach

Despite these advances, most approaches are embedded within a single monolithic end-to-end pipeline, which forces a trade-off: alignment is improved at the expense of preserving modality-specific richness. This reveals a fundamental limitation in current multimodal learning—the learning process itself, not just the fusion architecture, must be reconsidered [16]. Addressing this gap, we propose a decoupled paradigm that first builds a robust, fusion-optimized visual representation through a novel Mutual-Information Gated Distillation (MGD) strategy. We then freeze this representation before integrating auxiliary tabular data, ensuring that strong visual features are preserved and auxiliary information acts as a true complement rather than interference.

III. Method

We propose **MEDIATOR**, a decoupled learning framework for multimodal medical diagnosis that explicitly separates the representation learning of high-dimensional imaging data from the integration of auxiliary clinical information. The framework consists of two modules—(1) a **Vision Representation Module** dedicated to extracting fusion-optimized features from multiple imaging modalities, and (2) a **Final Classification Module** that incorporates structured patient records. Training proceeds in two distinct phases: Phase 1, where the Vision Representation Module is optimized using our Mutual-Information Gated Distillation (MGD) strategy, and Phase 2, where the Final Classification Module is trained with the vision module frozen.

A. *Overall Architecture*

The model architecture (Fig. 1) is designed to process two primary data streams: high-dimensional medical images (e.g., X-ray, CT) and low-dimensional tabular records (e.g., lab results, demographics).

- **Vision Representation Module**: Each imaging modality is encoded by a dedicated backbone network (e.g., ResNet, ViT), producing modality-specific features. These features are fused by a Vision Fusion Module, producing a unified representation. Phase 1 training of this module is guided by the MGD strategy, which balances modality contributions and aligns feature spaces.
- **Final Classification Module**: Structured records are encoded by a lightweight tabular MLP. A Final Fusion & Classifier block integrates the tabular features with visual features produced in Phase 1. Importantly, the Vision Representation Module remains frozen during Phase 2, preserving the visual representation while incorporating auxiliary information.

This architectural decoupling stabilizes optimization by isolating the representation learning from the subsequent multimodal integration stage.

B. *Phase 1: Vision Representation Learning via MGD*

The sole objective of Phase 1 is to learn a pure, high-quality visual representation optimized for fusion, without any influence from non-imaging data. To this end, we employ the **MGD (Mutual information Gating and Distillation)** methodology. MGD is fundamentally different from naive approaches that simply combine losses; it integrates dynamic, cooperative mechanisms directly into the learning process. This compels the model to autonomously discover a shared feature space that is inherently favorable for fusion, thereby addressing the core challenges of modality gap and representational dilution.

1) Feature Extraction and Fusion: The process begins with two independent, modality-specific vision backbones (e.g., ResNet, Vision Transformer) that encode their respective inputs:

$$F_x = \text{Encoder}_x(\text{Image}_x), \tag{1}$$

$$F_c = \text{Encoder}_c(\text{Image}_c), \tag{2}$$

where $F_x, F_c \in \mathbb{R}^D$ are the D-dimensional feature vectors extracted from the X-ray and CT images. These vectors are then merged by a **Vision Fusion Module**.

$$F_{fused} = \text{Fusion}(F_x, F_c). \tag{3}$$

The design of this module is a key experimental point; potential architectures range from simple concatenation to more sophisticated mechanisms like cross-attention, which can model complex inter-dependencies. The output is a fused feature, F_{fused}, which serves as the "teacher" in the subsequent distillation step.

2) Knowledge Distillation for Modality Gap Reduction: To reduce the modality gap, we distill the comprehensive knowledge encapsulated in the "teacher" feature, F_{fused}, back to the individual "student" features, F_x and F_c. This acts as a powerful regularization, forcing each backbone to align its feature space not for its own isolated task, but in a direction that is maximally beneficial for the final fusion. The distillation loss, $L_{distill}$, is formulated using the Mean Squared Error (MSE), which penalizes deviations between the student's feature and teacher's fused feature:

$$L_{distill} = \text{MSE}(F_x, \text{sg}(F_{fused})) + \text{MSE}(F_c, \text{sg}(F_{fused})). \tag{4}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. This operation is critical as it detaches the teacher from the computational graph of the students, ensuring that gradients flow only from the teacher to the students. This enforces a stable, unidirectional transfer of knowledge and prevents the teacher's representation from being corrupted by the students' learning process.

3) Information Gating and Feature Refinement: While knowledge distillation encourages learning shared, fusion-friendly features, the resulting representation (F_{fused}) may be dominated by the stronger modality, risking the suppression of unique, clinically vital information from the other. Information Gating is introduced as a critical mechanism to counteract this potential dominance and information loss. This mechanism employs small, learnable neural networks to compute a gate value for each modality, which dynamically identifies and controls the flow of the most salient original feature information.

$$\text{Gate}_x = \sigma(\text{Linear}_x(F_x)), \tag{5}$$

$$\text{Gate}_c = \sigma(\text{Linear}_c(F_c)), \tag{6}$$

where $\sigma(\cdot)$ is the Sigmoid function, constraining the gate values to the range [0, 1]. The final, refined feature, F_{final}, is then constructed by augmenting the fused feature with these gated original features. This additive, skip-connection-like step is crucial; it re-injects the most critical, modality-specific patterns that might have been diluted or lost during the initial fusion process that created F_{fused}. This allows the model to form a comprehensive representation that synergistically combines a holistic, fused view with crucial, preserved partial insights from each source.

$$F_{final} = F_{fused} + (\text{Gate}_x \odot F_x) + (\text{Gate}_c \odot F_c), \tag{7}$$

where $\odot$ denotes element-wise multiplication.

4) Loss Function for Phase 1: The total loss for Phase 1, L_{vision}, is a composite objective function designed to balance classification performance with feature space alignment:

$$L_{main} = \mathcal{L}_{CE}(\text{Classifier}(F_{final}), y), \tag{8}$$

where $\mathcal{L}_{CE}$ is the Cross-Entropy loss and y represents the ground-truth labels. The final loss function to be optimized is a weighted sum:

$$L_{vision} = L_{main} + \lambda \cdot L_{distill}. \tag{9}$$

The hyperparameter λ serves as a crucial trade-off coefficient, balancing the model's focus between the primary classification task (L_{main}) and the auxiliary regularization task of feature alignment ($L_{distill}$).

C. Phase 2: Final Classifier Training

Following Phase 1, the Vision Representation Module is **frozen** to preserve the learned visual embeddings and prevent catastrophic forgetting. This ensures that the visual representation, optimized for fusion among different imaging modalities, remains intact when introducing auxiliary modalities. Phase 2 is therefore restricted to training the final classifier that integrates structured patient data with the pre-trained visual features.

A lightweight Tabular MLP encodes the structured records:

$$F_{csv} = \text{TabularMLP}(\text{CSV data}). \tag{10}$$

where the encoder typically consists of a small stack of linear layers with ReLU activations and dropout.

The **Final Fusion & Classifier** block then integrates the tabular embedding (F_{csv}) with the visual features (F_x, F_c, F_{fused}). Different fusion strategies can be employed, ranging from simple concatenation to attention-based mechanisms that explicitly weigh the contribution of visual and tabular evidence. The classifier produces the final diagnostic prediction, optimized with the standard cross-entropy loss:

$$L_{final} = \mathcal{L}_{CE}(\text{FinalClassifier}(F_{visuals}, F_{csv}), y). \tag{11}$$

During this phase, gradients are propagated only through the Tabular MLP and the Final Fusion & Classifier, leaving the Vision Representation Module unchanged. This decoupled training protocol prevents overfitting to the simpler tabular features and enables a stable, effective integration of heterogeneous modalities.

IV. Experiments

We conduct a series of experiments to rigorously validate the proposed framework, MEDIATOR. Our experimental design follows a progressive structure: (1) identify the optimal unimodal encoder for each modality; (2) benchmark standard fusion architectures to establish strong baselines; (3) demonstrate the effectiveness of MEDIATOR through ablation studies and final multimodal evaluations.

A. Datasets and Implementation

We evaluate on a curated subset of the Stony Brook University COVID-19 Positive Cases collection [17], in accordance with the TCIA Data Usage Agreement. The subset comprises 1,500 paired imaging studies and clinical tables from 438 PCR-confirmed patients, each containing:

- **X-ray Image**: A frontal chest radiograph, predominantly portable anteroposterior (AP) views, converted from DICOM to PNG and resized to 224×224 pixels.
- **CT Image**: A representative axial slice from the thoracic section of a chest CT examination (either contrast-enhanced pulmonary angiography or routine non-contrast chest CT), also resized to 224×224 pixels.
- **Tabular Data**: 15 structured clinical features, with numerical features standardized and categorical ones one-hot encoded.

Since individual CT series of all study types for every patient contained over 50 of slices, dataset augmentation was performed by stride sampling: for each series we selected every k-th slice, where $k = \lceil N/50 \rceil$ for N original slices, producing up to 50 augmented CT slices per patient, each paired with the X-ray.

The diagnostic task is a 3-class severity diagnosis (mild, moderate, severe). We implement MEDIATOR in PyTorch with an NVIDIA A100 GPU, using AdamW (weight decay $1e-4$) and cosine annealing scheduler. The MGD hyperparameter λ was set to 0.4 for this experiment.

B. Backbone Selection for Unimodal Encoders

We first identify effective unimodal encoders for X-ray, CT, and tabular modalities. Table I reports unimodal accuracies. This selection is crucial, as the quality of the initial features directly impacts the potential of any subsequent fusion. We benchmarked several popular architectures for each modality independently, and the results are summarized in Table I.

TABLE I
PERFORMANCE OF UNIMODAL BACKBONE MODELS.

Modality	Backbone Model	pretrained	Accuracy(%)
X-ray	**Swin Transformer [18]**	SwinCheX [19]	**84.67**
	EfficientNet [21]	ImageNet [20]	82.33
	ResNet50 [22]	ImageNet [20]	81.00
	DenseNet121 [23]	ImageNet [20]	78.00
	DenseNet121 [23]	CheXpert [24]	77.00
CT	**3D ResNet18 [26]**	-	**56.33**
	3D ViT [27]	-	56.33
	3D ResNet50 [22]	Med3D [25]	48.33
	3D EfficientNet [28]	-	33.33
Tabular	CatBoost [29]	-	**84.48**
	RandomForest [30]	-	84.12
	ExtraTrees [31]	-	83.39
	MLP (Scikit-learn)	-	79.42
	Tabular MLP (Ours)	-	76.53

The unimodal results reveal a significant performance disparity among the modalities. The X-ray model is highly informative (up to 84.67% accuracy), while the CT (56.33%) and various tabular models (up to 84.48%) provide varied predictive signals. The results highlight the asymmetric predictive signal across modalities—X-ray and tabular features are highly informative, while CT is weaker. This asymmetry motivates the need for sophisticated fusion strategies. For subsequent experiments, we adopt the pretrained **Swin Transformer** for X-ray data, **ResNet18** for CT data, and our implemented **Tabular MLP** for the structured clinical records.

C. Analysis of Vision-Only Fusion

Having selected the optimal unimodal encoders, we next compared a range of standard vision-only fusion strategies to determine an effective integration scheme for X-ray and CT representations. This step is essential for identifying the most

suitable mechanism before introducing additional modalities. As summarized in Table II, several widely used fusion methods were evaluated, including concatenation, element-wise operations, and cross-attention variants.

TABLE II
PERFORMANCE OF BASELINE FUSION STRATEGIES

Fusion Strategy	Accuracy (%)
Unimodal Baselines	
X-ray Only (Swin Transformer)	84.67
CT Only (ResNet18)	56.33
Standard Vision-Only Fusion Baselines	
Concatenation	**87.33**
Element-wise Addition	85.67
Element-wise Product	86.33
Symmetry Cross Attention	86.33
Directional Cross Attention	84.00
Self-Attention (on Concat)	85.00

We observe that simple fusion methods provide only marginal gains over the stronger unimodal encoder (X-ray), and in some cases even degrade performance. Nevertheless, concatenation achieves the most stable improvements (87.33%) over the unimodal baselines, and thus we adopt this strategy as the fusion baseline in our subsequent MEDIATOR framework. This finding underscores the need for a more principled approach to address the asymmetry in predictive power between modalities.

D. Ablation Study of MEDIATOR

To validate the effectiveness of our proposed components, we conducted an ablation study on the MEDIATOR framework. The results, presented in Table III, demonstrate the contribution of each part of our two-phase strategy.

The initial vision-only MEDIATOR, without the MGD strategy's distillation loss ($\lambda = 0$), already shows competitive performance (85.33%). However, by enabling the full MGD strategy, the accuracy remarkably improves from 85.33% to **91.67%**, confirming that our distillation and gating mechanism is critical for creating a superior, fusion-optimized visual representation.

Next, we integrated the auxiliary tabular data. Adding tabular features with naive end-to-end fine-tuning improves to 96.67%. Crucially, **freezing the Phase 1 vision module** (our decoupled strategy) yields 98.33%, validating our hypothesis that frozen specialized encoders prevent representational dilution and enable superior integration.

TABLE III
ABLATION STUDY ON THE COMPONENTS OF MEDIATOR

Phase	Model / Configuration	Accuracy (%)
PHASE 1	MGD without distillation ($\lambda = 0$)	85.33
	MGD ($\lambda = 0.4$)	**91.67**
	MGD ($\lambda = 0.5$)	88.33
	MGD ($\lambda = 0.6$)	89.67
PHASE 2	MEDIATOR (End-to-End, Unfrozen)	96.67
	MEDIATOR (Frozen)	**98.33**

Overall, the ablation study of MEDIATOR confirms that each component—MGD distillation, structured fusion, and decoupled training—contributes to MEDIATOR's state-of-the-art performance.

V. CONCLUSION

We presented MEDIATOR, a novel decoupled two-phase framework designed to address core challenges in multimodal medical diagnosis. Our approach mitigates representational dilution by first constructing a robust, fusion-optimized visual representation via the proposed MGD strategy, and then freezing this module before integrating supplementary clinical data. Extensive experiments and ablation studies demonstrate that this decoupled paradigm consistently outperforms conventional end-to-end training, validating both the effectiveness of MGD and the importance of our two-phase design.

A key advantage of MEDIATOR lies in its flexibility and generality. While our experiments focused on fusing X-ray and CT images with structured clinical data, the framework is inherently modality-agnostic. The vision backbone can be readily extended to other forms of high-dimensional medical inputs such as MRI, histopathology slides, or non-visual modalities including genomics and clinical narratives.

Finally, the central principle underpinning our framework—decoupling representation learning from auxiliary data integration—extends beyond the medical domain. This paradigm provides a robust strategy for multimodal learning in settings where heterogeneous data with disparate dimensionalities must be fused, opening new avenues for cross-disciplinary applications of multimodal deep learning.

REFERENCES

[1] J. N. Acosta, C. Falcon, Y. J. Lee, and E. J. Topol, "Multimodal biomedical AI," *Nature Medicine*, vol. 28, no. 9, pp. 1773–1784, 2022.

[2] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Briefings in Bioinformatics*, Volume 23, Issue 2, bbab569, Mar. 2022.

[3] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. ACL*, 2019, pp. 6558–6569.

[4] S.-C. Huang, L.-S. D. Ku, and H.-H. Chen, "Fusion-based multimodal learning," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 1293–1301.

[5] Y. Zhang, H. Wang, J. Shi, C. Gao, H. C. Li, and G. Li, "MMC: Multimodal clinical data classification with curriculum learning," in *Proc. ICASSP*, 2022, pp. 1251–1255.

[6] S. Stahlschmidt, K. L. E. U. Ebneth, and H. P. A. van der Kooi, "Health-related multimodal machine learning: A review," *Frontiers in Big Data*, vol. 5, p. 868722, 2022.

[7] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597–1607.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.

[10] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," *arXiv preprint arXiv:2010.00747*, 2020.

[11] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proc. EMNLP*, 2023, pp. 1324–1336.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[13] D. Zhang, X. Han, and H. Deng, "MVAN: Multi-view attention network for multi-modal classification of Alzheimer's disease," in *Proc. BIBM*, 2020, pp. 713–718.
[14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
[15] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.
[16] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling Representation and Classifier for Long-Tailed Recognition," in *Proc. ICLR*, 2020.
[17] J. Saltz, M. Saltz, P. Prasanna, R. Moffitt, J. Hajagos, E. Bremer, J. Balsamo, and T. Kurc, "Stony Brook University COVID-19 Positive Cases," *The Cancer Imaging Archive*, 2021.
[18] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[19] A. Rosen, O. Tzemach, I. Gat, O. Geyer, I. Shelef, and I. Volf, "SwinCheX: A multi-scale feature learning and fusion network for chest X-ray diagnosis," *arXiv preprint arXiv:2206.04246*, 2022.
[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[21] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.
[23] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4700–4708.
[24] J. Irvin et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 590–597, 2019.
[25] B. Zheng, A. Gao, X. Huang, Y. Li, D. Liang, and X. Long, "A modified 3D EfficientNet for the classification of Alzheimer's disease using structural magnetic resonance images," *IET Image Processing*, vol. 17, no. 1, pp. 77–87, 2023.
[26] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proc. IEEE Int. Conf. Computer Vision Workshops* (ICCVW), 2017, pp. 3154-3160.
[27] Y. Wang, Z. Fan, T. Chen, H. Fan, and Z. Wang, "Can we solve 3D vision tasks starting from a 2D vision transformer?," *arXiv preprint*, Sep. 2022.
[28] Z. Huang, Y. Liu, and X. Tian, "Combine EfficientNet and CNN for 3D model classification," *Mathematical Biosciences and Engineering*, vol. 20, no. 5, pp. 5637-5657, 2023.
[29] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[31] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.