TGBD-Net: Tumor-Guided Bridging Distillation for Joint EGFR Mutation and Survival Prediction in Lung Cancer from CT Imaging

Huihui Yang, Shuo Wang (✉), Xingyu Huang, Haolin Sang, Chengcai Liu, Yi Wu, and Jie Tian

Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Engineering Medicine, Beihang University, Beijing, China
{23101068, shuo_wang}@buaa.edu.cn

Abstract. Epidermal growth factor receptor (EGFR) mutation prediction and progression-free survival (PFS) prediction are critical tasks for targeted therapy planning in lung cancer and can be jointly modeled from the same chest CT scan. In real-world clinical datasets, supervision is often inconsistent across patients: not all patients have labels for both tasks and tumor region annotations, leading to heterogeneous supervision across different patient subsets. However, most existing multi-task learning methods leverage a single shared representation across tasks, with limited consideration of representation consistency during training on these differently supervised patient subsets. In this study, we propose the Tumor-Guided Bridging Distillation Network (TGBD-Net), a unified framework for stabilizing representation learning under heterogeneous supervision, enabling joint EGFR mutation and PFS prediction. The proposed Tumor-Guided Feature Extraction (TGF) module regularizes feature learning using available tumor region annotations, while maintaining consistent behavior when such annotations are unavailable. In addition, a Bridging Distillation (BD) module enables cross-task knowledge transfer, allowing patients with single-task labels to contribute to joint representation learning. Experiments on a cohort of 7,348 patients with lung cancer demonstrate that TGBD-Net achieves an AUC of 0.823 for EGFR mutation prediction and a C-index of 0.743 for PFS prediction, highlighting its effectiveness in real clinical settings. Our code is available at `https://github.com/MenchYoung/TGBD-Net`.

Keywords: Lung cancer · Multi-task learning · CT imaging.

1 Introduction

EGFR mutation status [6,26] and progression-free survival (PFS) [5,21] are critical clinical endpoints for targeted therapy planning in lung cancer. Both tasks can be predicted from chest CT scans, which provide whole-lung context [7,23], and tumor region annotations, when available, provide auxiliary tumor-specific information. Given the clinical association between EGFR mutation status and

PFS [16], jointly modeling these endpoints, particularly with guidance from tumor annotations, may provide additional insights beyond separate prediction.

In this setting, multi-task learning (MTL) provides a natural paradigm for joint modeling across tasks [20, 25]. Most existing MTL methods primarily focus on optimization-level challenges, such as mitigating task conflicts [10, 13, 15, 19] and dynamic loss weighting [1, 3, 9], implicitly assuming that shared representations are learned under consistent supervision. However, effective joint modeling in real-world clinical datasets is challenged by heterogeneous supervision, characterized by: (1) not all patients have tumor region annotations, and (2) not all patients have both EGFR mutation labels and PFS follow-up information. As a result, the two tasks are supervised on different patient subsets. Under such heterogeneous supervision, MTL models built upon the shared backbone may learn task-dependent representations under different supervision types, leading to inconsistent representation learning.

Knowledge transfer approaches offer a potential solution by providing auxiliary supervision for label-missing samples. However, they typically restrict supervision to fully labeled samples to establish a teacher-student alignment [12, 17, 24]. As a result, their applicability under heterogeneous supervision is limited, as a large portion of partially labeled patients cannot be effectively utilized for cross-task knowledge transfer. Therefore, there is a need for a unified framework that simultaneously stabilizes representation learning and enables effective cross-task knowledge transfer under heterogeneous supervision.

To address these challenges, we propose the Tumor-Guided Bridging Distillation Network (TGBD-Net), a unified framework for stabilizing feature representation learning under heterogeneous supervision, which simultaneously supports EGFR mutation and PFS prediction. The main contributions are summarized as follows: (1) a unified framework that incorporates patients with single-task labels into joint learning; (2) two complementary representation-level mechanisms that regularize feature formation without requiring tumor annotations for all patients, and enable knowledge transfer across patient subsets with different task labels in a shared representation space; (3) extensive experiments on 7,348 patients demonstrating the robustness of TGBD-Net across multiple tasks.

2 Methodology

In this study, we propose TGBD-Net, a framework for joint EGFR mutation and PFS prediction under heterogeneous supervision. As illustrated in Fig. 1, given an input CT scan, an encoder together with a Tumor-Guided Feature Extraction (TGF) module outputs a shared feature $\mathbf{f}_{\text{shared}}$, which is fed into two task-specific heads to produce logits for EGFR mutation prediction task and a risk score for PFS prediction. The TGF module introduces an adaptive alignment mechanism to leverage additional tumor region annotations that are only available on a subset of patients. To stabilize shared representation learning under heterogeneous supervision, a Bridging Distillation (BD) module is employed to facilitate cross-task interactions via shared anchor features during training.

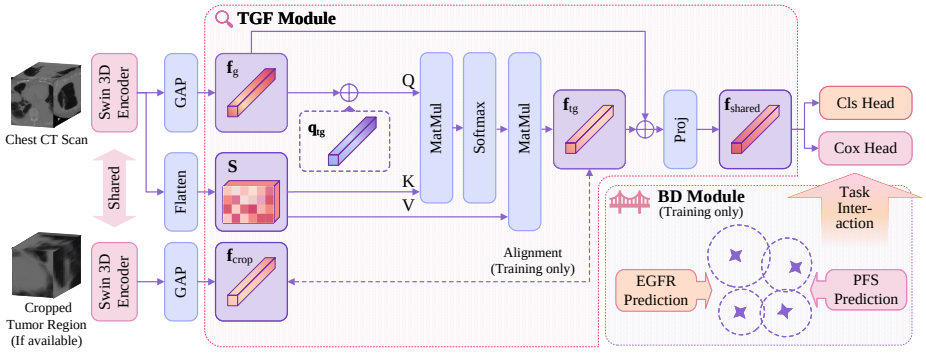


Fig. 1: Overview of TGBD-Net

2.1 The Tumor-Guided Feature Extraction Module

The Tumor-Guided Feature Extraction (TGF) module is designed to construct a tumor-guided shared representation from whole-lung CT inputs. Specifically, a Swin Transformer is adopted as the vision encoder to extract the feature map $\mathbf{F} \in \mathbb{R}^{C\times D\times H\times W}$ before the final normalization layer. A spatial token sequence $\mathbf{S} \in \mathbb{R}^{N\times C}$ $(N = D \times H \times W)$ is obtained by flattening $\mathbf{F}$, and a global feature $\mathbf{f}_{\mathrm{g}}$ is computed using global average pooling. To form an image-conditional query that adapts to each input, a learnable query token $\mathbf{q}_{\mathrm{tg}}$ is combined with $\mathbf{f}_{\mathrm{g}}$ to incorporate global information. The intermediate tumor-guided feature $\mathbf{f}_{\mathrm{tg}}$ is obtained via cross-attention and is subsequently fused with $\mathbf{f}_{\mathrm{g}}$ through concatenation and projection to form a shared feature $\mathbf{f}_{\mathrm{shared}}$ for downstream tasks:

$$\mathbf{f}_{\mathrm{tg}} = \mathrm{Attn}\big(\mathrm{query} = \mathbf{f}_{\mathrm{g}} + \mathbf{q}_{\mathrm{tg}},\ \mathrm{key, value} = \mathbf{S}\big), \tag{1}$$

$$\mathbf{f}_{\mathrm{shared}} = \mathrm{Proj}\Big(\mathrm{Concat}\big(\mathbf{f}_{\mathrm{g}},\ \mathbf{f}_{\mathrm{tg}}\big)\Big). \tag{2}$$

To adaptively leverage partial tumor region annotations (in the form of bounding boxes) available for only a subset of patients during training, while maintaining a unified modeling path for all patients, the TGF module introduces a conditional feature-level alignment strategy. For samples with available annotations, $\mathbf{f}_{\mathrm{tg}}$ is supervised to align with the cropped tumor feature $\mathbf{f}_{\mathrm{crop}}$, which is extracted using the same encoder and pooling strategy. For samples without annotations, $\mathbf{f}_{\mathrm{tg}}$ is generated without additional supervision. Through this strategy, the TGF module does not require tumor annotations for all samples, while enabling attention patterns learned from annotated samples to be implicitly shared with unannotated samples.

2.2 The Bridging Distillation Module

To enable cross-task knowledge transfer, the Bridging Distillation (BD) module mediates task-specific information through a set of K bridging anchors $\{\mathbf{c}_k\}_{k=1}^{K}$,

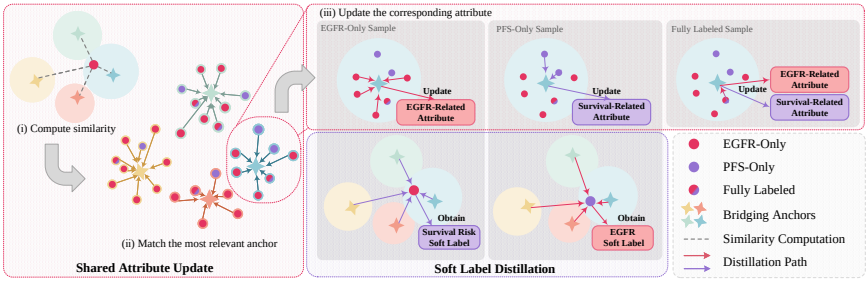


Fig. 2: Overview of the Bridging Distillation (BD) mechanism. For example, a sample labeled only with EGFR mutation updates the EGFR-related attribute and obtains a soft survival risk label through the shared anchors.

each equipped with learnable EGFR- and survival-related attributes. These anchors are initialized and periodically updated via K-means clustering, with the Hungarian algorithm applied to preserve anchor identities and their associated attributes. As illustrated in Fig. 2, BD operates in two stages:

(1) Shared Attribute Update. This stage transfers task-specific supervision to anchor-level attributes in a shared anchor space. For each task t (EGFR mutation and PFS prediction), only samples with labels for task t update the attribute of their most relevant anchor based on the cosine similarity between $\mathbf{f}_{\text{shared},i}$ obtained from the TGF module and each anchor:

$$m_k^{(t)} \leftarrow \alpha m_k^{(t)} + (1-\alpha)\frac{1}{|\mathcal{B}_k^{(t)}|}\sum_{i\in\mathcal{B}_k^{(t)}} y_i^{(t)}. \tag{3}$$

Here, $m_k^{(t)}$ denotes the anchor-level attribute for task t. $y_i^{(t)}$ refers to binary mutation labels for EGFR and survival risk scores from the Cox-supervised head for PFS. $\mathcal{B}_k^{(t)}$ denotes the labeled samples in the mini-batch assigned to anchor k for task t, and α is the momentum coefficient controlling the update rate.

(2) Soft Label Distillation. This stage distills anchor-level attribute values to samples lacking labels for task t to enable indirect supervision. The soft labels $\hat{y}_i^{(t)}$ are computed as a weighted combination of anchor-level attributes $m_k^{(t)}$ that are updated only from task-labeled samples:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{K} w_{ik}\, m_k^{(t)}, \quad w_{ik} = \frac{\exp(\text{sim}(\mathbf{f}_{\text{shared},i}, \mathbf{c}_k)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{f}_{\text{shared},i}, \mathbf{c}_j)/\tau)}. \tag{4}$$

Here, w_{ik} denotes the normalized cosine similarity between sample i and anchor k, with temperature τ controlling the smoothness of the assignment. $\hat{y}_i^{(t)}$ is used exclusively for the distillation loss and does not participate in inference.

2.3 Loss Functions

To accommodate heterogeneous supervision, the overall training objective combines task-wise completion losses and an auxiliary alignment constraint:

$$\mathcal{L} = \sum_t \underbrace{\left[\mathbb{I}_i^{(t)}\ \mathcal{L}_{\text{sup}}^{(t)} + (1 - \mathbb{I}_i^{(t)})\ \beta\ \mathcal{L}_{\text{BD}}^{(t)}\right]}_{\text{Task-wise Loss Completion}} + \gamma\ \mathbb{I}_{\text{anno},i}\ \mathcal{L}_{\text{align}}. \tag{5}$$

The loss incorporates task-specific switching via the indicator function $\mathbb{I}_i^{(t)}$, enabling consistent supervision with partially available labels. For samples labeled for task t ($\mathbb{I}_i^{(t)} = 1$), the supervised loss $\mathcal{L}_{\text{sup}}^{(t)}$ is applied, using binary cross-entropy (BCE) loss for EGFR mutation prediction and the Cox loss [4] for PFS prediction. For unlabeled samples ($\mathbb{I}_i^{(t)} = 0$), the bridging distillation loss $\mathcal{L}_{\text{BD}}^{(t)}$ provides auxiliary soft supervision by minimizing an MSE loss between model predictions and anchor-derived soft labels. To encourage feature-level consistency between the cropped tumor feature $\mathbf{f}_{\text{crop}}$ and the tumor-guided feature $\mathbf{f}_{\text{tg}}$, the alignment loss $\mathcal{L}_{\text{align}} = 1 - \cos(\mathbf{f}_{\text{tg}},\ \mathbf{f}_{\text{crop}})$ is applied exclusively to samples with available tumor region annotations ($\mathbb{I}_{\text{anno},i} = 1$), serving as a cosine distance penalty between the two representations.

3 Experiments and Results

3.1 Dataset and Preprocessing

This study included 7,348 patients with lung cancer from West China Hospital of Sichuan University, each with one chest CT scan acquired prior to treatment initiation and at least one task-specific ground-truth label. PFS was defined as the time from treatment initiation to disease progression or death. Among them, 6,484 patients had only EGFR mutation labels, 481 patients had only PFS follow-up information, and 383 patients had both labels, with 18.5% censored PFS cases, while 4,068 patients had available tumor region annotations. The dataset was split into training, validation, and test sets at the patient level with an 8:1:1 ratio, corresponding to 5,878, 735, and 735 patients, respectively. Patients with available PFS follow-up were stratified by progression time, while patients with only EGFR mutation labels were split independently. The final splits were formed by combining these mutually exclusive patient subsets. All lung volumes and annotated tumor regions, when available, were resized to a fixed resolution of $256 \times 256 \times 48$ voxels before being fed into the vision encoder.

3.2 Implementation Details

All models were implemented in PyTorch and trained on 2 NVIDIA GeForce RTX 4090 GPUs for 135 epochs with a batch size of 8. The AdamW optimizer was adopted with an initial learning rate of 10^{-4}, using a cosine learning rate schedule and linear warm-up. A weighted sampling strategy was used to increase

Table 1: Quantitative performance of different methods. – indicates that the method is not applicable to the corresponding task or evaluation metric. The p-values are obtained from the log-rank test on the Kaplan–Meier survival curves.

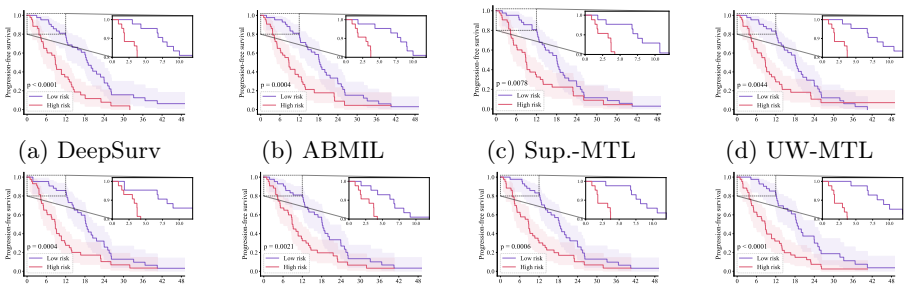
Sup.	Methods	AUC	Acc	Sens	Spec	C-index	p-value
Full	ResNet-50 [2]	0.6559	0.6217	0.5859	0.6545	-	-
	Swin Transformer [14]	0.7728	0.7164	0.6586	0.7705	-	-
	DeepSurv [11]	-	-	-	-	0.6584	<0.0001
Semi	Semi-SL (EMA) [22]	0.8091	0.7302	0.7393	0.7219	-	-
MIL	ABMIL [8]	0.8198	0.7493	**0.7485**	0.7500	0.6614	0.0004
Hetero.	Sup.-MTL [18]	0.8003	0.7317	0.6963	0.7640	0.6601	0.0078
	UW-MTL [3]	0.8068	0.7361	0.6748	**0.7921**	0.6829	0.0044
	CA-MTL [13]	0.7956	0.7111	0.6534	0.7640	0.6964	0.0004
	TP-MTL [9]	0.8043	0.7405	0.6963	0.7809	0.6908	0.0021
	JTR-MTL [15]	0.8019	0.7302	0.6871	0.7697	0.6772	0.0006
	TGBD-Net (Ours)	**0.8233**	**0.7522**	0.7362	0.7669	**0.7433**	<0.0001

the exposure of samples with PFS outcome during training. In addition, Bridging anchors ($K = 4$) were periodically updated every 5 epochs from epoch 10 to 100. The hyperparameters were set to $\alpha = 0.9$ for Eq. (3), $\tau = 0.2$ for Eq. (4), and $\beta = 0.5$, $\gamma = 1.0$ for Eq. (5). The best-performing model was selected based on validation AUC and the same model was used for evaluating PFS prediction.

3.3 Comparison Results

We compare our method with representative approaches across four supervision types (Sup.): (1) Full Supervision, including a ResNet-50 initialized with MedicalNet pretraining weights [2], a Swin Transformer [14], and a Cox-based DeepSurv [11]; (2) Semi-supervised learning, represented by an EMA-based semi-supervised strategy (Semi-SL.) [22]; (3) Multiple Instance Learning (MIL), represented by ABMIL [8]; (4) Heterogeneous Supervision (Hetero.) multi-task learning, including standard supervised multi-task learning (Sup.-MTL) [18] as well as MTL variants with uncertainty-based weighting, conflict-averse gradient descent, task prioritization, and joint task regularization (UW-MTL [3], CA-MTL [13], TP-MTL [9], and JTR-MTL [15]).

To ensure a fair comparison, all baselines (except ResNet-50) adopt the same Swin Transformer backbone as TGBD-Net, while ABMIL and all MTL baselines are trained on the full dataset. Furthermore, ResNet-50 and Swin Transformer are included as reference backbones and performance anchors, while DeepSurv is inherently limited to survival analysis and does not support EGFR mutation prediction. The semi-SL. baseline is evaluated only on the classification task, as the label for PFS involves both progression time and event status, which is not directly compatible with EMA-based distillation. Since ABMIL is not originally designed for 3D CT-based models, we adapt it by computing attention-based aggregation over visual tokens extracted from 3D volumes.



(a) DeepSurv (b) ABMIL (c) Sup.-MTL (d) UW-MTL

(e) CA-MTL (f) TP-MTL (g) JTR-MTL (h) TGBD-Net

Fig. 3: Kaplan–Meier survival curves for different methods, with a zoomed-in inset of the early follow-up period; shaded areas show the 95% confidence intervals.

(1) Quantitative Comparison. As shown in Table 1, TGBD-Net achieves the best overall performance on both EGFR mutation prediction and PFS prediction. **For EGFR mutation prediction,** we evaluate model performance using the area under the ROC curve (AUC), accuracy (Acc), sensitivity (Sens), and specificity (Spec). TGBD-Net achieves the highest AUC (0.823) and Acc (0.752) among all compared methods, while maintaining the best overall balance between Sens and Spec. **For PFS prediction,** the concordance index (C-index) is used as the evaluation metric. The p-values reported in Table 1, Table 2, and Fig. 3 are obtained from log-rank tests on Kaplan–Meier curves, where patients are stratified into high- and low-risk groups based on the median predicted risk scores of each model. TGBD-Net achieves the highest C-index of 0.743, representing an improvement of 0.047 over the second-best CA-MTL model and 0.082 over the ABMIL baseline, which performs competitively on EGFR mutation prediction. Compared with optimization-level MTL methods, the proposed BD module performs representation-level regularization through bridging anchors, leading to more stable cross-task learning under heterogeneous supervision.

(2) Kaplan–Meier Analysis. Fig. 3 shows the Kaplan–Meier survival curves for different methods. Among the compared methods, TGBD-Net demonstrates clearer risk stratification of high- and low-risk groups, particularly in the early follow-up period, with a log-rank p-value < 0.0001.

3.4 Ablation Studies

We conduct three ablation studies to validate TGBD-Net's architecture, key design choices, and training stability.

(1) TGF and BD Modules. Table 2 reports ablations of different combinations of TGF and BD modules, where BD-reg, BD-uni, and Full denote different BD variants. Specifically, BD-reg only incorporates anchor-based regularization without distillation, while BD-uni adds uni-directional distillation using bridging anchors with only EGFR-related attribute. Compared with Sup.-MTL, adding

Table 2: Effectiveness of tumor-guided feature learning strategy and relevant modules, with Sup.-MTL as the baseline. Bold values indicate the best result.

Variant	TGF	BD variant	AUC	Acc	C-index	p-value
Sup.-MTL	✗	✗	0.8003	0.7317	0.6601	0.0078
+ TGF	✓	✗	0.8154	0.7449	0.6539	0.0238
+ TGF + BD-reg	✓	anchor regularization	0.8078	0.7185	0.6905	0.0002
+ TGF + BD-uni	✓	uni-directional distillation	0.8220	**0.7537**	0.6768	<0.0001
Full (Ours)	✓	bi-directional distillation	**0.8233**	0.7522	**0.7433**	<0.0001

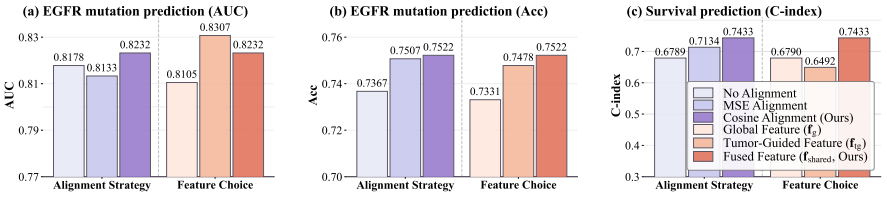


Fig. 4: Ablation of alignment strategies and feature choices. (a) AUC and (b) Accuracy (Acc) for EGFR mutation prediction; (c) C-index for PFS prediction.

TGF yields higher EGFR mutation prediction (AUC/Acc: 0.815/0.745), while the C-index remains comparable (0.654). Incorporating the BD module shows an overall improving trend across tasks. BD-reg mainly enhances PFS prediction by providing anchor-based regularization, while BD-uni favors EGFR mutation prediction through uni-directional knowledge transfer, reflecting different cross-task regularization behaviors across BD variants. The full bi-directional BD yields a 0.089 gain in C-index over the TGF-only model, while maintaining competitive performance in EGFR mutation prediction, highlighting the importance of coordinated cross-task interactions under heterogeneous supervision.

(2) Alignment Strategies and Feature Choices. Fig. 4 ablates alignment strategies and feature choices for EGFR (AUC/Acc) and PFS (C-index). For EGFR mutation prediction, while the MSE-based alignment shows less stable behavior, the cosine similarity-based alignment yields consistent gains, with AUC from 0.818 to 0.823 and Acc from 0.737 to 0.752. Across different features chosen for the EGFR prediction head, both the tumor-guided feature $\mathbf{f}_{\text{tg}}$ and the shared feature $\mathbf{f}_{\text{shared}}$ outperform the global feature $\mathbf{f}_{\text{g}}$, with $\mathbf{f}_{\text{tg}}$ achieving the highest AUC (0.831). However, for PFS prediction, $\mathbf{f}_{\text{shared}}$ yields a C-index of 0.743 that notably surpasses $\mathbf{f}_{\text{tg}}$ (0.649). Therefore, $\mathbf{f}_{\text{shared}}$ is selected as the final representation for both tasks to achieve a balanced performance.

(3) Anchor Parameterization Strategies. Compared with our periodic update strategy in the BD module, using a fixed K-means clustering after warm-up without updates results in inferior performance (AUC: 0.816, C-index: 0.701), while treating anchors as fully learnable parameters optimized via gradient descent further degrades performance (AUC: 0.803, C-index: 0.645).

4 Conclusion

In this study, we propose TGBD-Net, a unified framework for joint EGFR mutation and PFS prediction from chest CT scans under heterogeneous supervision. By leveraging limited tumor region annotations through tumor-guided feature alignment and structuring cross-task interactions via bridging distillation, the proposed method enables more consistent representation learning by incorporating samples with missing labels and without annotations. Extensive experiments demonstrate that TGBD-Net achieves favorable performance across both tasks, highlighting its robustness in realistic clinical settings. Future work will further explore external multicenter validation and broader clinical generalization under heterogeneous supervision.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 62576022), the National Key R&D Program of China under Grant No. 2023YFC2506502, Beijing Natural Science Foundation (L232131), and the Fundamental Research Funds for the Central Universities.

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

References

1. Bytez.com, Qin, X., Wang, X., Liao, N., Yan, J.: NTKMTL: Mitigating Task Imbalance in Multi-Task Learning... (Dec 2025), https://bytez.com/docs/neurips/119235/paper
2. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
3. Cipolla, R., Gal, Y., Kendall, A.: Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. pp. 7482–7491 (Jun 2018). https://doi.org/10.1109/CVPR.2018.00781
4. D. R. Cox: Regression Models and Life-Tables. The Journal of the Royal Statistical Society, Series B (Statistical Methodology) **34**(2), 187–202 (Jan 1972). https://doi.org/10.1111/j.2517-6161.1972.tb00899.x
5. Du, X., Lu, X., Cao, X.: Gantt chart for updated OS and PFS after cancer targeted therapy. The Innovation Medicine **1**, 1 (Jan 2023). https://doi.org/10.59717/j.xinn-med.2023.100008
6. Gregory J. Riely, Douglas E. Wood, Dara L. Aisner, Billy W. Loo Jr, Andrea L. Axtell, Jessica R. Bauman, Lisa Hang: NCCN Guidelines® Insights: Non–Small Cell Lung Cancer, Version 7.2025. Journal of the National Comprehensive Cancer Network **Volume 23**(Issue 9) (Sep 2025). https://doi.org/10.6004/jnccn.2025.0043
7. Huang, X., Wang, S., Liu, C., Sang, H., Wu, Y., Tian, J.: Feature Copy-Paste Network for Lung Cancer EGFR Mutation Status Prediction in CT images . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15974. Springer Nature Switzerland (September 2025)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based Deep Multiple Instance Learning. In: Proceedings of the 35th International Conference on Machine Learning. pp. 2127–2136 (Jul 2018), iSSN:

9. Jeong, W., Yoon, K.J.: Quantifying Task Priority for Multi-Task Optimization. pp. 363–372 (Jun 2024). https://doi.org/10.1109/CVPR52733.2024.00042
10. Jeong, W., Yoon, K.J.: Selective task group updates for multi-task optimization (2025), https://arxiv.org/abs/2502.11986
11. Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology **18** (Feb 2018). https://doi.org/10.1186/s12874-018-0482-1
12. Kim, D., Shim, H.: Classifier-guided clip distillation for unsupervised multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4661–4671 (June 2025)
13. Liu, B., Liu, X., Jin, X., Stone, P., Liu, Q.: Conflict-Averse Gradient Descent for Multi-task Learning. In: Advances in Neural Information Processing Systems. vol. 34 (2021)
14. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video Swin Transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3192–3201. IEEE, New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.00320
15. Nishi, K., Kim, J., Li, W., Pfister, H.: Joint-Task Regularization for Partially Labeled Multi-Task Learning. pp. 16152–16162 (Jun 2024). https://doi.org/10.1109/CVPR52733.2024.01529
16. Nishimura, T., Fujimoto, H.: PFS, OS or toxicity: what is the most important factor in the treatment of EGFR-mutated lung cancer? Annals of Oncology **36**(2), 220–221 (Feb 2025). https://doi.org/10.1016/j.annonc.2024.10.010
17. Parchami-Araghi, A., Böhle, M., Rao, S., Schiele, B.: Good Teachers Explain: Explanation-Enhanced Knowledge Distillation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. vol. 15131, pp. 293–310. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-73464-9_18, series Title: Lecture Notes in Computer Science
18. Ruder, S.: An Overview of Multi-Task Learning in Deep Neural Networks (Jun 2017). https://doi.org/10.48550/arXiv.1706.05098
19. Senushkin, D., Patakin, N., Kuznetsov, A., Konushin, A.: Independent Component Alignment for Multi-Task Learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20083–20093. IEEE, Vancouver, BC, Canada (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.01923
20. Shen, J., Zhen, X., Wang, Q., Worring, M.: Episodic multi-task learning with heterogeneous neural processes. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 75214–75228. Curran Associates, Inc. (2023)
21. Stewart, D.J., Ramsay, T., Navani, V., Liu, G., Jiang, D.M., Batist, G.: Progression-Free Survival Gain: A Reliable Primary End Point for Drug Registration That Can Accelerate Patient Access to Urgently Needed Therapies. Journal of Clinical Oncology **42**(8), 973–974 (Mar 2024). https://doi.org/10.1200/JCO.23.02259
22. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. pp. 1195–1204. Red Hook, NY, USA (Dec 2017)
23. Yan, J., Han, X., Zhao, J., Gao, Y.: Multimodal Hypergraph Guide Learning for Non-Invasive ccRCC Survival Prediction . In: proceedings of Medical Image

Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15971. Springer Nature Switzerland (September 2025)
24. Zhang, Q., Hao, X., Chen, Q., Xu, R., Cong, F., Lu, C., Xu, H.: Multi-modal Knowledge Decomposition based Online Distillation for Biomarker Prediction in Breast Cancer Histopathology . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15974. Springer Nature Switzerland (September 2025)
25. Zhang, Y., Yang, Q.: A Survey on Multi-Task Learning. IEEE Transactions on Knowledge and Data Engineering **34**(12), 5586–5609 (Dec 2022). https://doi.org/10.1109/TKDE.2021.3070203
26. Zhou, F., Guo, H., Xia, Y., Le, X., Tan, D.S.W., Ramalingam, S.S., Zhou, C.: The changing treatment landscape of EGFR-mutant non-small-cell lung cancer. Nature Reviews Clinical Oncology **22**(2), 95–116 (Feb 2025). https://doi.org/10.1038/s41571-024-00971-2