Confidence-Aware Supervision for Robust Multi-Task Embryo Grading

Bogyu Park[1], Yongwon Jo[1], Jiyeon Kang[1], Jin Heo[1], and Hyung Min Kim[1(✉)]

Kai Health, Seoul, Republic of Korea
{bogyu.park,yongwon.jo,jiyeon.kang}@kaihealth.tech
jinh2720@gmail.com
hikhm0304@naver.com (✉)

Abstract. Multi-task learning is widely adopted for automated embryo grading. However, the reliability of predicted confidence across heterogeneous grading components remains underexplored. We observed that blastocyst development stage, inner cell mass, and trophectoderm exhibit distinct uncertainty characteristics, and that overconfident predictions may become more evident when outputs are jointly evaluated. To improve robustness under heterogeneous uncertainty, we propose a confidence-aware soft-label learning framework that selectively refines supervision based on feature consistency and entropy stability. Unlike uniform regularization, the proposed approach preserves discriminative structure while reducing excessive confidence in unstable spaces. Experiments on 28,670 Day-5 embryo images from seven IVF clinics showed that our method achieves the highest overall mF1 (0.7218) while reducing overconfident error rates relative to baseline training. When combined with temperature scaling, calibration further improves without performance degradation. These findings highlight that uncertainty-aware supervision provides a practical strategy for enhancing reliability in multi-task medical AI systems.

Keywords: Confidence-aware learning · Multi-task learning · Embryo grading.

1 Introduction

Embryo selection in in vitro fertilization (IVF) relies on morphological assessment based on the Gardner grading (GG) system [1], which evaluates developmental stage (Stage), inner cell mass (ICM), and trophectoderm (TE). Despite its widespread clinical use, the grading system remains subjective and shows substantial inter-observer variability [2], motivating the development of automated deep learning systems.

Recent studies adopt multi-task learning (MTL) to jointly predict Stage, ICM, and TE [3, 4]. While shared representations enable strong classification performance, these grading components are inherently heterogeneous: Stage reflects global morphology, whereas ICM and TE depend on localized structures

with greater intrinsic ambiguity. Consequently, confidence characteristics differ across tasks.

Most training strategies, however, apply uniform supervision across samples and tasks. Such indiscriminate optimization may produce unstable confidence behavior, particularly for ambiguous cases. In this work, we analyze reliability patterns in multi-task embryo grading and propose a confidence-aware soft-label learning framework that selectively refines stable predictions while preserving ambiguous instances.

Experiments on 28,670 Day-5 embryo images from seven IVF clinics show that our framework improves robustness and reduces overconfident errors without sacrificing classification performance. Moreover, it complements post-hoc calibration techniques such as temperature scaling, which operate at inference time, by improving representation stability.

2 Related Works

Deep Learning for Embryo Grading. Convolutional neural networks-based approaches have been widely adopted to predict Stage and blastocyst quality [5, 6]. Recent work employs MTL to jointly predict Stage, ICM, and TE [3, 4], achieving strong classification performance but with limited analysis of confidence behavior.

Calibration in Medical AI. Metrics such as expected calibration error (ECE) [7, 8] are commonly used to assess confidence reliability [9, 10]. Methods including label smoothing[11, 12] and temperature scaling [13] reduce overconfidence but are typically studied in single-task settings with uniform regularization.

Multi-Task Learning under Heterogeneous Uncertainty. MTL tasks often exhibit varying difficulty and uncertainty profiles [14, 15]. Prior work focuses on adaptive loss weighting for performance balancing [16], while confidence stability remains less explored. In embryo grading, structural differences between Stage reflecting global morphology and ICM & TE relying on localized structures motivate supervision strategies that account for heterogeneous uncertainty.

3 Methods

3.1 Problem Formulation

Given an embryo image x_i, the model predicts probability distributions $\hat{y}_i^{\text{Stage}}$, $\hat{y}_i^{\text{ICM}}$, and $\hat{y}_i^{\text{TE}}$ for the three grading tasks. Standard MTL optimizes per-task cross-entropy losses with equal weighting:

$$\mathcal{L}_{\text{CE}} = \sum_{t \in \{\text{Stage, ICM, TE}\}} \text{CE}(y_i^t, \hat{y}_i^t). \tag{1}$$

Task-specific uncertainty heterogeneity results in differing levels of confidence stability.

3.2 Confidence-Aware Soft Label Refinement

To enhance representation stability while avoiding noisy supervision, we refine GG labels based on feature consistency and cluster purity. We first extract temporally smoothed features using parameters updated via exponential moving average (EMA) [17]:

$$h_i = f_{\theta_{\text{EMA}}}(x_i). \tag{2}$$

The features $\{h_i\}$ are grouped into K clusters corresponding to the number of classes for each task ($K = 5$ for Stage and $K = 3$ for ICM & TE). For cluster k, the empirical label distribution p_k is computed from ground-truth labels within each cluster, and the purity is defined as

$$\gamma_k = \max_c p_{k,c}. \tag{3}$$

Sample-level confidence combines cluster purity and feature consistency:

$$\alpha_i = \text{clip}\big(\gamma_{k(i)} \cdot \cos(h_i, \mu_{k(i)}), \alpha_{\min}, \alpha_{\max}\big), \tag{4}$$

where $\mu_{k(i)}$ is the cluster centroid, $\alpha_{\min} = 0.2$, and $\alpha_{\max} = 0.9$.

Let α_q denote the q-th percentile of confidence scores. Only samples in the top-$(1-q)$ percentile are refined:

$$y_i = \begin{cases} \alpha_i p_{k(i)} + (1-\alpha_i) y_i^{\text{orig}}, & \alpha_i \geq \alpha_q, \\ y_i^{\text{orig}}, & \text{otherwise.} \end{cases} \tag{5}$$

This selective mechanism preserves the original hard labels for ambiguous samples while refining supervision for high-confidence samples. By refining structurally consistent and confident instances, the model can stabilize class prototypes without amplifying noise from uncertain spaces.

3.3 Entropy-Guided Update Scheduling

To avoid premature refinement, we monitor predictive entropy during training. Predictive entropy is defined as $H(p) = -\sum_c p_c \log p_c$, and averaged across samples at each epoch to obtain E_t. We compute the moving average entropy variation:

$$\bar{E}_t = \frac{1}{M} \sum_{j=t-M+1}^{t} |E_j - E_{j-1}| \tag{6}$$

and measure its relative change $\Delta E_t = \bar{E}_t / E_0$, where E_0 denotes the reference entropy level.

Soft label updates are triggered when $\Delta E_t < \tau$ for L consecutive epochs. After each update, the reference entropy is reset ($E_0 \leftarrow E_t$), ensuring that subsequent refinements are evaluated relative to the most recently stabilized state.

3.4 Training Objective

Each task is optimized using either original hard labels or refined soft labels:

$$\mathcal{L}_{\text{task}} = \begin{cases} \text{CE}(\mathbf{y}_i^{\text{orig}}, \hat{\mathbf{y}}_i), & \text{hard label,} \\ \text{KL}(\mathbf{y}_i, \hat{\mathbf{y}}_i), & \text{soft label.} \end{cases} \tag{7}$$

The overall loss is

$$\mathcal{L}_{\text{total}} = \sum_{\text{task}} \mathcal{L}_{\text{task}}. \tag{8}$$

3.5 Post-hoc Calibration

To isolate the effect of feature-level refinement from inference-time adjustment, we additionally apply temperature scaling (TS) at test time. Given logits z, calibrated probabilities are

$$\hat{y}^{\text{TS}} = \text{Softmax}\left(\frac{z}{T}\right), \tag{9}$$

where temperature T is optimized on the validation set by minimizing negative log-likelihood. TS is a standard post-hoc calibration method and is not a contribution of this work; it is included to show that representation-level refinement and inference-time calibration act as complementary mechanisms.

3.6 Hyperparameter Selection

We perform a grid search over the entropy threshold τ and the maximum number of update steps. The final configuration is selected based on validation performance, considering both aggregated classification accuracy and calibration metrics.

4 Experimental Results

4.1 Dataset and Setup

Dataset. We use 28,670 Day-5 embryo images collected from seven IVF clinics (2012–2020) under IRB approval. Images were annotated by consensus of three embryologists following the GG system. The dataset was split into training (16,888), validation (5,667), and test (6,115) sets without patient overlap. It includes five Stage classes and three classes each for ICM and TE. The dataset cannot be publicly released due to IRB and patient-privacy constraints; it is a multi-center heterogeneous cohort from seven clinics rather than a single-site dataset.

Training Details. We employed ResNet18 [18] as the backbone architecture. The models were trained using AdamW (lr=1e-4) with a 5-epoch warmup and cosine decay over 100 epochs. Experiments were repeated five times with different random seeds. We compared baseline training with hard labels, label smoothing ($\epsilon = 0.1$), and the proposed method. Temperature scaling (TS) was additionally applied at inference time using validation data to assess post-hoc calibration effects.

Soft Label and Update Settings. For confidence-aware refinement, features h_i (128-d) were obtained via a learned linear projection layer from the 512-d backbone representation. Task-specific K-means clustering (with k-means++ initialization) was then applied to these features ($K = 5$ for Stage and $K = 3$ for ICM & TE) and recomputed at each refinement stage using EMA features. The top 50% of samples ranked by confidence were considered eligible for soft-label updates. Focusing on high-confidence and structurally consistent samples helps stabilize class prototypes while preventing noisy refinement of intrinsically ambiguous instances.

Entropy-guided scheduling employed a moving window of $M = 3$ epochs and a patience of $L = 3$ epochs. For entropy thresholds $\tau \in \{0.01, 0.02, 0.03\}$, we performed a grid search over maximum update counts ranging from 3 to 7. For each τ, the best configuration was selected based on validation performance considering overall average macro-F1 (mF1) and reliability metrics. Among these candidates, the final model reported in the main results corresponds to the configuration that achieves the highest aggregated mF1 while preserving improved reliability relative to the baseline ($\tau = 0.01$, 3 updates). Post-hoc temperature scaling was applied to the trained model without altering learned representations. The exponential moving average decay was set to 0.999. Behavior was relatively stable across the explored τ values and update counts, and the overhead remains modest since refinement is applied only a few times during training.

4.2 Evaluation Metrics

We evaluate classification performance using mF1 together with reliability-related metrics.

ECE [7, 8] measures the alignment between predicted confidence and empirical accuracy:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|,$$

where B_m denotes the m-th confidence bin ($m = 10$).

Overconfident Error Rate (OER) [13] quantifies the proportion of incorrect predictions made with high confidence ($\sigma = 0.8$):

$$\mathrm{OER} = \frac{1}{N} \sum_{i=1}^{N} I(\hat{y}_i \neq y_i \wedge \max_c p_{i,c} \geq \sigma).$$

Table 1. Aggregated performance comparison under joint-task evaluation. Comparison of baseline training, label smoothing (LS), post-hoc temperature scaling (TS), and the proposed confidence-aware soft-label learning method (3-0.01). Metrics are computed by pooling predictions across tasks. **Bold**/underline indicate best/second-best results.

Method	Avg. mF1 ↑	ECE ↓	Entropy ↑	OER ↓
Baseline	$\underline{0.7206 \pm 0.0046}$	0.0599 ± 0.0243	1.1953 ± 0.0051	0.0857 ± 0.0201
Baseline + TS	0.7206 ± 0.0046	$\mathbf{0.0099 \pm 0.0023}$	1.2093 ± 0.0014	0.0405 ± 0.0036
Label Smoothing	0.6913 ± 0.0047	0.0365 ± 0.0109	$\mathbf{1.2235 \pm 0.0021}$	$\mathbf{0.0185 \pm 0.0046}$
Label Smoothing + TS	0.6913 ± 0.0047	0.0116 ± 0.0028	$\underline{1.2175 \pm 0.0012}$	$\underline{0.0304 \pm 0.0037}$
Proposed (3-0.01)	$\mathbf{0.7218 \pm 0.0053}$	0.0292 ± 0.0148	1.2039 ± 0.0062	0.0574 ± 0.0190
Proposed (3-0.01) + TS	0.7218 ± 0.0053	$\underline{0.0113 \pm 0.0051}$	1.2082 ± 0.0009	0.0420 ± 0.0016

Prediction Entropy measures predictive uncertainty,

$$H(p_i) = -\sum_{c=1}^{C} p_{i,c} \log p_{i,c},$$

and is used to characterize confidence dispersion.

Entropy Gap (E-Gap) [19] evaluates uncertainty separation between correct and incorrect predictions:

$$\text{E-Gap} = E[H(p) \mid y \neq \hat{y}] - E[H(p) \mid y = \hat{y}].$$

Higher E-Gap indicates better uncertainty discrimination.

System-Level Aggregation. To examine reliability when outputs are jointly considered, we pool predictions from all three tasks and compute Entropy, ECE, and OER over the combined set of $3N$ outputs. This aggregation provides a complementary view of confidence behavior that may not be fully captured by averaging per-task metrics. In contrast, average mF1 is computed as the arithmetic mean of the three task-wise mF1 scores.

E-Gap is reported at the task level to evaluate uncertainty differentiation between correct and incorrect predictions. When predictions from heterogeneous tasks are pooled, this separation becomes less directly interpretable; therefore, we report average entropy to summarize overall uncertainty dispersion.

4.3 Joint-Task Aggregated Performance

Table 1 summarizes aggregated metrics obtained by pooling predictions across tasks. The baseline achieves competitive classification accuracy but exhibits a measurable proportion of overconfident errors under joint evaluation. Uniform label smoothing reduces overconfident errors, yet this improvement is accompanied by a noticeable decline in classification performance. In contrast, the

Table 2. Task-wise performance and reliability metrics. Classification and reliability metrics for Stage, ICM, and TE prediction reported separately for each task.

2*Method	mF1 ↑			ECE ↓		
	Stage	ICM	TE	Stage	ICM	TE
Baseline	0.8245 ± 0.0065	**0.6154 ± 0.0153**	0.7220 ± 0.0072	0.0345 ± 0.0171	0.0748 ± 0.0300	0.0723 ± 0.0262
Baseline + TS	0.8245 ± 0.0065	0.6154 ± 0.0153	0.7220 ± 0.0072	0.0113 ± 0.0028	0.0188 ± 0.0046	**0.0134 ± 0.0020**
LS	0.8220 ± 0.0094	0.5870 ± 0.0087	0.6649 ± 0.0167	0.0493 ± 0.0092	0.0192 ± 0.0107	0.0509 ± 0.0121
LS + TS	0.8220 ± 0.0094	0.5870 ± 0.0087	0.6649 ± 0.0167	0.0156 ± 0.0043	**0.0124 ± 0.0035**	0.0203 ± 0.0036
Proposed (3-0.01)	**0.8273 ± 0.0086**	0.6137 ± 0.0128	**0.7244 ± 0.0038**	0.0191 ± 0.0081	0.0368 ± 0.0158	0.0437 ± 0.0133
Proposed (3-0.01) + TS	0.8273 ± 0.0086	0.6137 ± 0.0128	0.7244 ± 0.0038	**0.0094 ± 0.0017**	0.0168 ± 0.0069	0.0187 ± 0.0072

2*Method	OER ↓			E-Gap ↑		
	Stage	ICM	TE	Stage	ICM	TE
Baseline	0.0533 ± 0.0118	0.1014 ± 0.0278	0.1024 ± 0.0218	0.3887 ± 0.0265	**0.0948 ± 0.0079**	**0.1250 ± 0.0052**
Baseline + TS	0.0345 ± 0.0046	0.0382 ± 0.0041	0.0418 ± 0.0036	**0.4268 ± 0.0091**	0.0669 ± 0.0039	0.0943 ± 0.0056
LS	**0.0276 ± 0.0042**	**0.0151 ± 0.0062**	**0.0129 ± 0.0045**	0.3589 ± 0.0224	0.0395 ± 0.0057	0.0452 ± 0.0074
LS + TS	0.0438 ± 0.0053	0.0191 ± 0.0070	0.0370 ± 0.0074	0.3768 ± 0.0226	0.0431 ± 0.0077	0.0620 ± 0.0084
Proposed (3-0.01)	0.0427 ± 0.0061	0.0627 ± 0.0224	0.0669 ± 0.0295	0.3935 ± 0.0217	0.0774 ± 0.0144	0.1114 ± 0.0062
Proposed (3-0.01) + TS	0.0352 ± 0.0017	0.0439 ± 0.0039	0.0484 ± 0.0056	0.4068 ± 0.0307	0.0681 ± 0.0057	0.1029 ± 0.0149

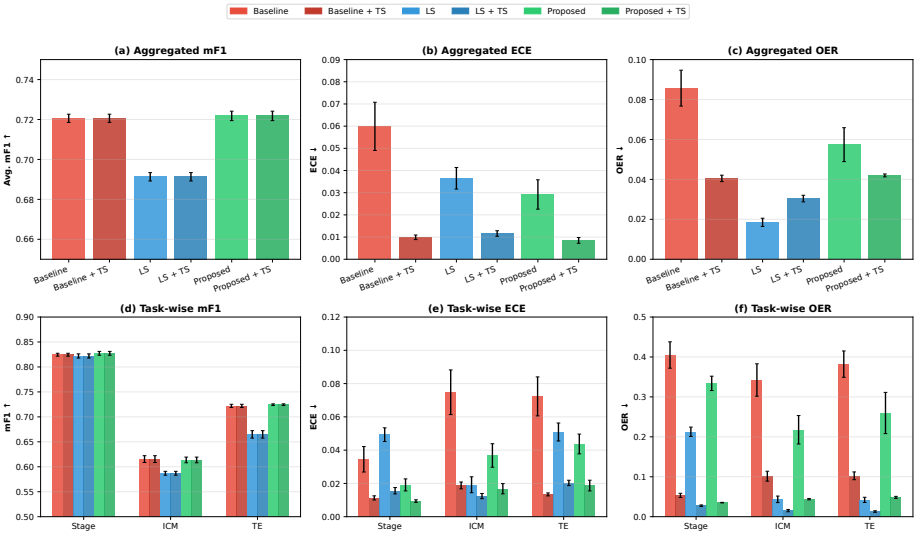

Fig. 1. Aggregated and task-wise performance comparison. The top row presents metrics computed under joint-task aggregation (Avg. mF1, ECE, OER), while the bottom row reports task-wise results for Stage, ICM, and TE. Temperature scaling (TS) consistently reduces calibration error across configurations. The proposed method preserves classification performance while lowering overconfident error rates relative to the baseline. Error bars denote standard error over five runs.

proposed configuration (3–0.01) maintains the highest average mF1 while reducing overconfident errors compared to the baseline, without sacrificing classification performance. When temperature scaling is applied to the same trained model, calibration improves further without affecting accuracy, indicating that representation-level refinement and post-hoc adjustment provide complementary benefits.

4.4 Task-wise Analysis

Table 2 reports task-wise metrics. Across all methods, uncertainty differentiation is consistently higher for Stage than for ICM and TE, reflecting differences in structural complexity. The effect of uniform smoothing varies across grading components. Although label smoothing reduces overconfident errors, it is accompanied by a noticeable decline in classification performance, particularly for ICM and TE. In contrast, the proposed method preserves Stage performance while reducing overconfident errors for ICM and TE relative to the baseline.

Although the E-Gap decreases compared to the baseline, it remains consistently higher than that of label smoothing, suggesting that the proposed approach retains a greater degree of uncertainty differentiation.

4.5 Comparison between Task-wise and Aggregated Evaluation

Figure 1 compares task-wise metrics with those computed under joint-task aggregation. While per-task ECE values appear moderate, aggregated evaluation reveals a higher concentration of overconfident errors in the baseline. The proposed method reduces these aggregated overconfident errors relative to uniform label smoothing, while maintaining competitive classification performance.

4.6 Discussion

Biological Perspective. Confidence differences reflect inherent grading characteristics: Stage involves global structural transitions, whereas ICM and TE rely on finer cellular features with greater intrinsic ambiguity. This supports the presence of heterogeneous uncertainty across tasks.

Clinical Relevance. Post-hoc calibration adjusts confidence at inference time, while uniform label smoothing may suppress discriminative structure. The proposed selective refinement provides a training-time alternative that preserves accuracy while attenuating excessive confidence in unstable regions.

Limitations. Uncertainty separation for ICM and TE remains limited, reflecting intrinsic ambiguity. Future work may explore task-specific supervision strategies and longitudinal modeling.

5 Conclusion

We examined reliability characteristics in multi-task embryo grading and demonstrated that heterogeneous grading components exhibit distinct confidence patterns. Our analysis shows that evaluation under joint-task aggregation offers a complementary reliability perspective beyond per-task metrics. To address heterogeneous uncertainty during training, we introduced a confidence-aware soft-label learning framework that selectively refines stable predictions while preserving ambiguous instances. Our method maintains competitive performances while reducing overconfident errors under aggregated evaluation.

When combined with temperature scaling, calibration improvements can be obtained without performance degradation, indicating that representation-level refinement and inference-time adjustment play complementary roles. Overall, this study highlights the importance of uncertainty-aware supervision in heterogeneous multi-task medical AI systems and provides a practical framework for improving reliability without sacrificing discriminative performance.

Acknowledgments. This work was supported by the Korea Medical Device Development Fund by the Korea government (Ministry of Science and ICT, Ministry of Trade, Industry and Energy, Ministry of Health & Welfare, Ministry of Food and Drug Safety) [grant number RS-2023-00254175].

Ethical Approval. This retrospective study using de-identified Day-5 embryo images was approved by the Institutional Review Boards (IRB) of all seven participating IVF clinics (approval numbers: GMH-202301, 2024-RESEARCH-1, MMIRB2024-6, RTR-2023-01, B-2208-772-104, HIIRB2023-01, and 2204-003-113).

Disclosure of Interests. All authors are employed by Kai Health, which develops the technology evaluated in this study. The authors have no other competing interests to declare.

References

1. David K Gardner, William B Schoolcraft, et al. In vitro culture of human blastocysts. *Towards reproductive certainty: fertility and genetics beyond*, pages 378–388, 1999.
2. Hyung Min Kim, Hyoeun Kang, Chaeyoon Lee, Jong Hyuk Park, Mi Kyung Chung, Miran Kim, Na Young Kim, and Hye Jun Lee. Evaluation of the clinical efficacy and trust in ai-assisted embryo ranking: survey-based prospective study. *Journal of Medical Internet Research*, 26:e52637, 2024.
3. Lazaros Moysis, Lazaros Alexios Iliadis, George Vergos, Sotirios P Sotiroudis, Achilles D Boursianis, Achilleas Papatheodorou, Konstantinos-Iraklis D Kokkinidis, Mohammad Abdul Matin, Panagiotis Sarigiannidis, Ilias Siniosoglou, et al. Artificial intelligence-empowered embryo selection for ivf applications: A methodological review. *Machine Learning and Knowledge Extraction*, 7(2):56, 2025.
4. Alessia Auriemma Citarella, Pietro Battistoni, Chiara Coscarelli, Fabiola De Marco, Luigi Di Biasi, and Mengyuan Wang. Embryovision ai: An explainable deep learning framework for enhanced blastocyst selection in assisted reproductive technologies. *Image and Vision Computing*, page 105795, 2025.
5. Mahdi-Reza Borna, Mohammad Mehdi Sepehri, and Behnam Maleki. An artificial intelligence algorithm to select most viable embryos considering current process in ivf labs. *Frontiers in Artificial Intelligence*, 7:1375474, 2024.
6. Jacob Theilgaard Lassen, Mikkel Fly Kragh, Jens Rimestad, Martin Nygård Johansen, and Jørgen Berntsen. Development and validation of deep learning based embryo selection across multiple days of transfer. *Scientific Reports*, 13(1):4235, 2023.
7. Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

8. Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.
9. Yuanhang Zheng, Zeshui Xu, Tong Wu, and Zhang Yi. A systematic survey of fuzzy deep learning for uncertain medical data. *Artificial Intelligence Review*, 57(9):230, 2024.
10. Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878, 2020.
11. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
12. Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, 2019.
13. Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
14. Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
15. Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.
16. SM Kamrul Hasan and Cristian A Linte. A multi-task cross-task learning architecture for ad hoc uncertainty estimation in 3d cardiac mri image segmentation. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4. IEEE, 2021.
17. Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
18. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
19. Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.