Concept-Consistent Semi-Supervised Learning for Concept Bottleneck Models
via Confidence-Guided Pseudo-Label Propagation

Anonymous Authors

Anonymous Institution

Abstract. Concept Bottleneck Models (CBMs) improve interpretability by routing predictions through human-understandable concepts, but their practical use is limited by the high cost of concept annotation. Existing semi-supervised CBM methods address this issue with pseudo-label transfer and heatmap-based spatial alignment, but these designs introduce architectural complexity and rely on intermediate spatial feature maps. In this paper, we propose **FixCBM**, a semi-supervised framework that learns from unlabeled data through confidence-guided consistency in concept space. **FixCBM** generates pseudo-labels from weakly augmented views, filters them by per-concept confidence, and enforces consistency on strongly augmented views. A warmup schedule further stabilizes training by delaying unlabeled supervision until pseudo-label quality becomes sufficiently reliable. Experiments on four benchmark datasets show that **FixCBM** consistently improves both concept accuracy and task accuracy over prior semi-supervised CBM baselines. With only 10% labeled data, **FixCBM** surpasses fully supervised CBM and CEM baselines in concept accuracy on CUB-200-2011, and outperforms SSCBM across all four datasets. Additional analyses, including labeled-ratio evaluation and test-time intervention, show that **FixCBM** learns reliable and interventionable concept representations. These results suggest that enforcing reliable consistency directly in concept space provides a simpler and more effective alternative to spatial alignment for semi-supervised concept learning, leading to more robust performance.

Keywords: Concept Bottleneck Models · Semi-Supervised Learning · Pseudo-Labeling · Consistency Learning · Concept-Level Consistency

1 Introduction

Deep neural networks achieve strong performance in visual recognition, yet their opacity limits use in high-stakes domains such as medical diagnosis [18], autonomous systems [2], and legal reasoning [7]. Concept Bottleneck Models (CBMs) [11] address this by introducing a human-interpretable bottleneck, where semantic concepts are predicted from the input and used to infer the final label. This design enables a decision chain and supports test-time intervention, allowing users to correct concept predictions and observe the resulting output.

Despite their appeal, CBMs require costly concept annotations for every image, which is prohibitive at scale (e.g., CUB-200-2011). Some works bypass annotation using language models [13,23], but rely on GPT-scale priors and lack reliable evaluation of extracted concepts.

This setting is important yet lacks a standard solution. Existing methods either remove concept supervision, sacrificing grounded evaluation, or rely on architecture-dependent pseudo-supervision that fails to ensure reliable concept predictions. Moreover, standard semi-supervised methods focus on class labels, while CBMs require reliable supervision over a multi-label concept space with varying confidence.

SSCBM [9] uses KNN transfer and spatial alignment for unlabeled data, but the latter introduces architectural constraints, scales poorly with concept number, and is sensitive to noisy early embeddings. This suggests the core challenge lies in reliable concept-level supervision rather than spatial alignment. Early noisy pseudo-labels can destabilize learning, motivating approaches grounded in concept semantics and robust to uncertainty.

Our key insight is that semi-supervised concept learning should focus on concept-level prediction reliability rather than feature-level spatial alignment. If a concept predictor is confident under weak augmentation, its prediction should remain consistent under stronger perturbations. This yields a concept-level supervisory signal that is naturally compatible with multi-label prediction and robust to early uncertainty by ignoring low-confidence concepts.

To this end, we propose **FixCBM**, a semi-supervised framework that replaces spatial alignment with weak-to-strong consistency in concept space. For each unlabeled image, pseudo-labels are generated from a weak view, filtered by per-concept confidence masking, and enforced on a strong view. A warmup schedule further stabilizes training by delaying the effect of unlabeled supervision.

This design has three advantages: it is spatial-feature-free, requiring no intermediate representations; concept-aware, evaluating pseudo-label reliability per concept; and training-stable, where confidence masking and warmup suppress noisy pseudo-labels and reduce confirmation bias.

Our contributions are threefold:

- We identify reliable concept-level supervision on unlabeled data as the key challenge in semi-supervised CBMs, and show that weak-to-strong consistency in concept space provides a direct and effective alternative to spatial alignment.
- We propose **FixCBM**, a semi-supervised framework that combines concept-space pseudo-labeling, per-concept confidence masking, and warmup-based consistency scheduling, without requiring intermediate spatial feature maps or architecture-specific modifications.
- We show empirically on four benchmarks that this formulation improves both concept and task accuracy over prior semi-supervised CBM baselines, while preserving intervention quality and concept faithfulness. These results suggest that explicit spatial alignment is not strictly necessary for overall effective semi-supervised concept learning.

2 Related Work

Concept Bottleneck Models. constrain predictions to pass through human-interpretable concepts. Concept Embedding Models (CEMs) [6] replace scalar concept probabilities with vector embeddings, improving flexibility and downstream performance. IntCBM [3] improves the use of user interventions during training. Label-free CBM [13] and Post-hoc CBM [23] remove annotation requirements via language model priors, but sacrifice grounded evaluation and direct concept supervision.

Semi-supervised Learning for CBMs. SSCBM [9] is the most related work, using KNN pseudo-labeling and concept heatmap alignment. In contrast, our method enforces prediction consistency across augmented views without relying on intermediate spatial representations. Related work has also explored weakly supervised concept learning, but not the semi-supervised CBM setting with a consistency-based objective.

Semi-supervised Learning with Consistency Regularization. Consistency regularization [14,12] enforces stable predictions under perturbations. Mean Teacher [17], MixMatch [1], FixMatch [15], and UDA [22] are representative methods. Unlike these class-level approaches, FixCBM applies the same principle to multi-label concept prediction, where each concept must be calibrated independently.

Several adjacent lines of work are also relevant. Strong augmentation (e.g., RandAugment [5]) supports the weak-to-strong scheme used in FixCBM. Adaptive thresholding methods such as FlexMatch [24] and SoftMatch [4] improve pseudo-label selection and could extend to per-concept settings. Related work on semi-supervised multi-label classification [16] addresses label-wise calibration but does not consider interpretability or concept-level constraints.

3 Method

We adopt the standard Concept Embedding Model (CEMs) [6], in which concept predictions are first produced and then mapped to the task label. In the semi-supervised setting, the training data are divided into a labeled set D_L, containing both class and concept labels, and an unlabeled set D_U, containing only class labels. We use A_{weak} and A_{strong} to denote weak and strong augmentation functions, respectively, and write $p(x) \in [0,1]^k$ for the vector of per-concept predicted probabilities produced by the model on input x.

We present **FixCBM**, a semi-supervised framework for Concept Embedding Models that achieves reliable concept learning under scarce annotation by enforcing prediction consistency in concept probability space. Figure 1 gives an overview of our framework. We describe each component below.

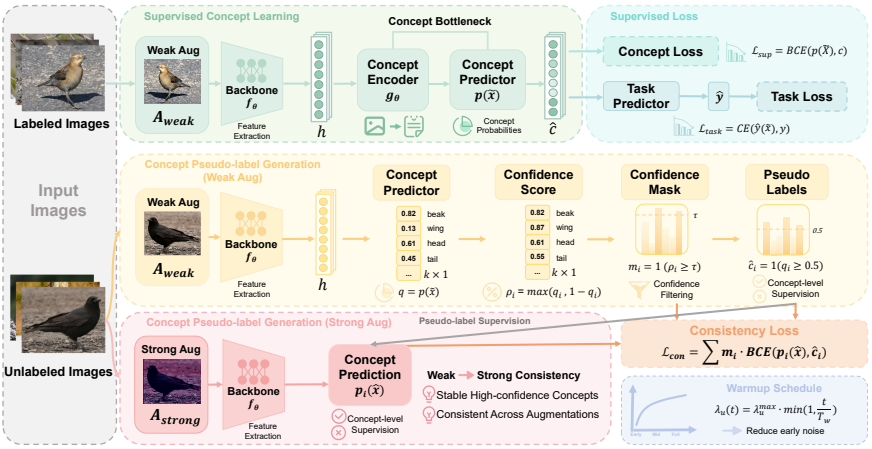

Fig. 1. Framework overview of **FixCBM**. The model learns from labeled data via supervised losses, and effectively leverages unlabeled data through weak-to-strong consistency with confidence masking and warmup scheduling.

3.1 Overview and Motivation

The core idea of **FixCBM** is to enforce weak-to-strong consistency in concept space. If a weakly augmented view yields a high-confidence prediction for a concept, then the same concept prediction should remain stable under a stronger augmentation, providing a direct supervisory signal for unlabeled data without requiring spatial alignment or intermediate spatial feature maps.

The key difference from SSCBM [9] is illustrated in Figure 1. SSCBM generates pseudo-labels via KNN in feature space and enforces alignment between concept embeddings and spatial heatmaps; both operations depend on global feature retrieval or intermediate spatial features.

In contrast, **FixCBM** filters pseudo-labels by per-concept prediction confidence and enforces consistency directly at the output of the concept predictor, yielding a simpler and more direct formulation.

3.2 Labeled Data: Concept Embedding Learning

For labeled samples $(x, y, \boldsymbol{c}) \in \mathcal{D}_L$, we first apply weak augmentation $\tilde{x} = A_{\text{weak}}(x)$ and then feed $\tilde{x}$ into the CEM backbone to obtain concept probabilities $\boldsymbol{p}(\tilde{x})$ and corresponding task prediction $\hat{y}(\tilde{x})$.

The **labeled concept loss** supervises the concept bottleneck with ground-truth concept labels:

$$\mathcal{L}_{\text{sup}} = \frac{1}{|\mathcal{B}_L|} \sum_{(x,y,\boldsymbol{c}) \in \mathcal{B}_L} \text{BCE}\big(\boldsymbol{p}(\tilde{x}),\, \boldsymbol{c}\big), \tag{1}$$

where $\mathcal{B}_L$ is the labeled mini-batch and BCE denotes binary cross-entropy applied independently per concept. Let $\mathcal{B}_U \subseteq \mathcal{D}_U$ denote the unlabeled mini-batch, and let $\mathcal{B} = \mathcal{B}_L \cup \mathcal{B}_U$ be the full mini-batch.

The **task loss** is applied to all samples (both labeled and unlabeled), since class labels are available for all data:

$$\mathcal{L}_{\text{task}} = \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \text{CE}\big(\hat{y}(\tilde{x}),\, y\big), \tag{2}$$

where $\hat{y}(\tilde{x})$ denotes the task prediction from the weakly-augmented view.

3.3 Unlabeled Data: Confidence-Guided Concept Consistency

For each unlabeled sample $x \in \mathcal{D}_U$, we generate two augmented views: a weak view $\tilde{x} = \mathcal{A}_{\text{weak}}(x)$ and a strong view $\hat{x} = \mathcal{A}_{\text{strong}}(x)$.The weak view produces concept pseudo-labels, while the strong view is used for consistency training.

Pseudo-label Generation. We first compute concept probabilities on the weakly augmented view,

$$\boldsymbol{q} = \boldsymbol{p}(\tilde{x}) = (q_1, \ldots, q_k) \in [0,1]^k. \tag{3}$$

and convert them into hard pseudo-labels by thresholding at 0.5:

$$\hat{c}_i = 1[q_i \geq 0.5]. \tag{4}$$

Confidence Masking. Not all pseudo-labels are equally trustworthy. We retain only concepts whose predictions are confidently above or below the decision boundary. For concept i, define the confidence as $\rho_i = \max(q_i, 1 - q_i)$. The confidence mask is:

$$m_i = 1[\rho_i \geq \tau], \tag{5}$$

where $\tau \in (0, 1)$ is a confidence threshold tuned on the validation set.

Consistency Loss. We apply the unsupervised loss on the strongly augmented view, using only the retained pseudo-labels from the weak view as supervision. pseudo-labels:

$$\mathcal{L}_{\text{con}} = \frac{1}{|\mathcal{B}_U|} \sum_{x\in\mathcal{B}_U} \frac{1}{k} \sum_{i=1}^{k} m_i \cdot \text{BCE}\big(p_i(\hat{x}),\, \hat{c}_i\big), \tag{6}$$

where $\mathcal{B}_U$ is the unlabeled mini-batch and $p_i(\hat{x})$ is the predicted probability for concept i on the strongly-augmented view. Note that the pseudo-label $\boldsymbol{q}$ is detached from the computation graph, so no gradient flows through it.

3.4 Warmup Schedule for Stable Training

Early in training, concept embeddings are poorly initialized and pseudo-labels on unlabeled data are unreliable. Applying the full consistency loss from the first epoch can destabilize the labeled-data concept learning. To address this, we introduce a linear warmup schedule for the consistency loss weight:

$$\lambda_u(t) = \lambda_u^{\max} \cdot \min\left(1, \frac{t}{T_w}\right), \tag{7}$$

where t is the current training epoch, T_w is the warmup duration (default $T_w = 10$ epochs), and $\lambda_u^{\max}$ is the maximum unlabeled loss weight. This allows the model to first learn a minimally stable concept representation from labeled supervision before the full unlabeled consistency signal is applied.

Empirical Validation of the Warmup Design. Figure 3 shows that the mask rate increases steadily during training, indicating improving pseudo-label quality. This validates the role of warmup in preventing unreliable early pseudo-labels from degrading training. This trend clearly supports the warmup design: applying the full $\lambda_u^{\max}$ weight when mask rates are very low would force the model to learn from noisy, low-confidence pseudo-labels, significantly corrupting the already established representation by the labeled-data loss.

3.5 Final Objective

The overall training objective of **FixCBM** is:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_s \, \mathcal{L}_{\text{sup}} + \lambda_u(t) \, \mathcal{L}_{\text{con}} \tag{8}$$

where $\mathcal{L}_{\text{task}}$ is the task classification loss, $\mathcal{L}_{\text{sup}}$ is the supervised concept loss on labeled data, and $\mathcal{L}_{\text{con}}$ is the confidence-guided consistency loss on unlabeled data. Here, λ_s denotes the weight of the supervised concept loss, and $\lambda_u(t)$ is the warmup-modulated weight of the unlabeled consistency loss.

4 Experiments

We evaluate **FixCBM** on four benchmark datasets across four axes: (1) comparison with fully supervised upper bounds, (2) concept and task accuracy under varying labeled ratios, (3) test-time intervention, and (4) ablation study.

4.1 Experimental Setup

Datasets. We evaluate on four real-world fine-grained image recognition benchmarks: **CUB-200-2011** [20] (200 bird species, 312 binary concepts, 11,788 images), **AwA2** [21] (50 animal classes, 85 continuous attributes, 37,322 images), **PBC** [19] (5 white blood cell types, concept-annotated), and **7-point** [10] (skin lesion diagnosis, 7 clinical criteria as concepts). These datasets span a variety of concept types (binary/continuous), domain expertise levels, and class imbalance characteristics, providing testbed for semi-supervised concept learning.

Evaluation Metrics. We report concept accuracy (C-Acc) and task accuracy (Y-Acc). Concept accuracy is computed by averaging prediction accuracy across all concepts, while task accuracy is measured as standard classification accuracy. For datasets with continuous attributes (e.g., AwA2), attributes are binarized following SSCBM for consistent evaluation.

Baselines. We compare **FixCBM** with the following baselines:

- **CBM** [11]: vanilla Concept Bottleneck Model in the fully supervised setting, serving as a basic reference.
- **CEM** [6]: Concept Embedding Model in the fully supervised setting, our primary fully-supervised upper bound.
- **CEM-SS**: CEM trained with k-NN pseudo-labels only (no alignment or consistency loss), isolating the benefit of our consistency objective.
- **SSCBM** [9]: the state-of-the-art semi-supervised CBM, using k-NN pseudo-labels and heatmap alignment loss.
- **Label-free CBM** [13]: unsupervised CBM relying on LLM-generated concept sets, included for reference.

All semi-supervised baselines consistently use the same labeled ratio r and the same train/val/test split across all experiments, ensuring strict fair comparison.

Implementation Details. All models use ResNet-34 [8] pretrained on ImageNet as the backbone, following SSCBM. The concept context generator uses LeakyReLU activations with embedding dimension $m = 64$. We train all models using SGD with momentum 0.88, weight decay 5×10^{-6}, and learning rate 0.016, with ReduceLROnPlateau scheduling. We train for 70 epochs with batch size 32 and check validation every 5 epochs. Weak augmentation consists of random horizontal flip and crop; strong augmentation uses RandAugment [5] ($N = 2$, $M = 10$). The confidence threshold is $\tau = 0.876$ (tuned via Optuna on CUB; see Appendix for dataset-specific values). Warmup duration is $T_w = 10$ epochs. All experiments are run with 3 independent random seeds. All experiments are conducted on a single NVIDIA RTX 4090 GPU with standard settings.

4.2 Main Results: Concept and Task Accuracy

Table 1 reports concept accuracy (C-Acc) and task accuracy (Y-Acc) on all four datasets at a labeled ratio of $r = 0.1$. **FixCBM** consistently outperforms all semi-supervised baselines across both metrics. On CUB-200-2011, it achieves 91.57% concept accuracy and 75.92% task accuracy, improving over SSCBM by +1.52% and +9.14%, respectively. On AwA2, **FixCBM** achieves 97.43% concept accuracy and 90.57% task accuracy, surpassing SSCBM and even slightly exceeding the fully supervised CEM baseline in task accuracy, demonstrating that our consistency loss acts as an effective regularizer. Furthermore, the gap between CEM-SS and **FixCBM** highlights our contribution, with gains of +24.03% on PBC and +8.95% on 7-point in concept accuracy.

Table 1. Concept accuracy (C-Acc, %) and task accuracy (Y-Acc, %) comparison on four benchmarks at labeled ratio $r = 0.1$. **Bold**: best semi-supervised result. †: fully-supervised upper bound. ‡: unsupervised (no concept labels used). Results averaged over 3 seeds.

Method	CUB-200-2011		AwA2		PBC		7-point	
	C-Acc	Y-Acc	C-Acc	Y-Acc	C-Acc	Y-Acc	C-Acc	Y-Acc
CBM† [11]	88.95	41.47	78.96	90.03	89.31	99.58	55.50	55.44
CEM† [6]	92.27	69.00	91.29	90.17	87.07	99.81	55.48	62.53
Label-free CBM‡ [13]	–	74.52	–	68.72	–	36.31	–	48.78
CEM-SS (KNN only)	83.16	58.84	66.19	87.73	70.91	99.61	65.37	63.54
SSCBM [9]	90.05	66.78	95.67	88.56	94.05	99.61	69.62	66.58
FixCBM (Ours)	**91.57**	**75.92**	**97.43**	**90.57**	**94.94**	**99.74**	**74.32**	**67.09**

Table 2. Effect of labeled ratio r on CUB-200-2011. C-Acc and Y-Acc (%) reported. K=1 denotes a 1-shot-per-class setting. Results averaged over 3 seeds.

Method	K=1		r=0.05		r=0.10		r=0.15		r=0.20	
	C	Y	C	Y	C	Y	C	Y	C	Y
CEM-SS	81.27	45.96	82.64	52.00	84.11	58.42	83.70	60.99	84.87	61.60
SSCBM [9]	88.13	58.27	90.62	68.93	90.05	66.78	91.32	69.33	92.01	71.06
FixCBM (Ours)	**88.19**	**68.86**	**90.62**	**75.39**	**91.57**	**75.92**	**91.95**	**75.40**	**92.71**	**75.70**
CEM† (full sup.)	92.27 / 69.00		92.27 / 69.00		92.27 / 69.00		92.27 / 69.00		92.27 / 69.00	

4.3 Effect of Labeled Ratio

Table 2 reports the performance on CUB-200-2011 under varying labeled ratios, including the extremely low-label K=1 setting and ratios from $r = 0.05$ to $r = 0.20$. Overall, **FixCBM** consistently outperforms SSCBM across all settings.

The advantage of **FixCBM** is particularly pronounced in the low-supervision regime. Under the K=1 setting, **FixCBM** improves task accuracy from 58.27% to 68.86%, achieving a substantial gain of +10.59 percentage points. At $r = 0.05$, the gain remains significant, with an improvement of +6.46 points. Even at higher labeled ratios, **FixCBM** continues to outperform SSCBM, demonstrating its robustness across different levels of supervision.

This trend suggests that our confidence-guided consistency is more robust to pseudo-label noise than heatmap-based alignment. In low-label regimes where pseudo-label quality is inherently limited, filtering unreliable predictions and enforcing weak-to-strong consistency provides a more stable supervisory signal, leading to improved concept and task performance.

4.4 Test-Time Intervention

Figure 2 shows task accuracy as a function of the number of intervened concept groups. Models with better concept representations exhibit steeper, monotonically increasing curves under intervention.

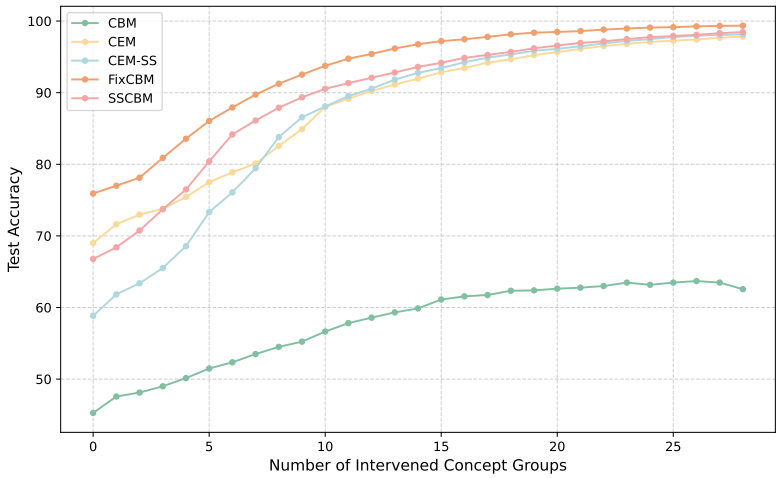

Fig. 2. Test-time intervention results. Task accuracy improves as more concept labels are corrected. **FixCBM** achieves the steepest improvement curve among semi-supervised methods, indicating high-quality concept representations with strong intervention capability.

FixCBM achieves the best intervention performance among semi-supervised models, outperforming SSCBM and semi-supervised CEM, especially in the low-intervention regime, indicating more reliable concept predictions.

As more concepts are corrected, all models converge to similar high accuracy, suggesting that residual errors mainly stem from imperfect concept predictions. The fully supervised CBM is included as a reference: although its accuracy is lower due to the strict bottleneck, it notably shows consistent improvement under intervention, further confirming its interpretability.

4.5 Ablation Study

Table 3 reports ablations on the CUB-200-2011 dataset at $r = 0.1$. We study the following design choices:

1. **w/o consistency loss** ($\lambda_u = 0$): removes unlabeled supervision entirely, equivalent to supervised CEM on labeled data only.
2. **w/o confidence masking** ($\tau = 0$, all pseudo-labels retained): removes the quality gate, using all pseudo-labels regardless of confidence.
3. **w/o warmup** ($T_w = 0$): applies the full consistency loss from epoch 1.
4. **w/o strong augmentation**: replaces $\mathcal{A}_{\text{strong}}$ with $\mathcal{A}_{\text{weak}}$, testing whether augmentation asymmetry is necessary.
5. **Full FixCBM**: the complete model.

Table 3. Ablation study on CUB-200-2011 (r=0.1). Each row removes one component of **FixCBM**. C-Acc / Y-Acc (%) reported.

Configuration	C-Acc	Y-Acc
w/o consistency loss ($\lambda_u = 0$)	90.65	69.38
w/o confidence masking ($\tau = 0$)	90.32	72.51
w/o warmup ($T_w = 0$)	90.62	71.57
w/o strong augmentation	90.62	70.87
FixCBM (full model)	**91.57**	**75.92**

Key findings: Removing the consistency loss leads to the largest performance drop, reducing task accuracy from 75.92% to 69.38%, confirming that unlabeled consistency learning is the primary source of improvement. Removing confidence masking degrades performance, lowering task accuracy to 72.51%, showing that filtering unreliable pseudo-labels is crucial for stable learning. Removing warmup further reduces performance to 71.57%, indicating that introducing unlabeled supervision too early harms training stability. Without strong augmentation, task accuracy drops to 70.87%, demonstrating that the weak-to-strong consistency gap is essential for effective regularization.

4.6 Analysis: Pseudo-label Quality and Warmup Dynamics

Mask rate over training. Figure 3 plots the pseudo-label mask rate on CUB-200-2011 ($r = 0.1$). With warmup ($T_w = 10$), the mask rate starts low, increases steadily (to around 55% by epoch 10), and stabilizes at 70%–75%, indicating improving pseudo-label quality. Without warmup ($T_w = 0$), it starts higher but remains below the warmup curve and saturates at a lower level, suggesting that early unreliable pseudo-labels limit final performance. Overall, warmup improves pseudo-label quality by delaying unlabeled supervision, while confidence masking filters unreliable predictions.

Multi-label consistency vs. single-label FixMatch. Applying FixMatch to the task label is suboptimal for two reasons: concept prediction requires per-concept confidence calibration, while a single threshold discards entire samples; moreover, consistency over k concepts provides richer supervision than task-level learning. These differences explain the effectiveness of **FixCBM**.

4.7 Computational Cost Comparison

Table 4 compares the computational cost of different methods on CUB-200-2011 ($r = 0.1$). Compared with SSCBM, **FixCBM** requires slightly more training time per epoch, mainly due to the additional weak/strong forward passes for unlabeled samples. However, **FixCBM** is significantly more memory-efficient, reducing GPU memory usage from 14.29 GB to 5.13 GB. This is because our

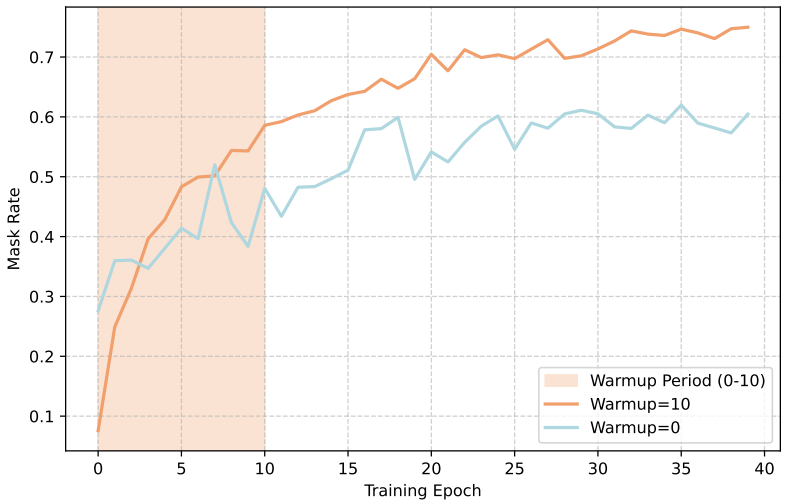

Fig. 3. Pseudo-label mask rate on CUB-200-2011 ($r = 0.1$). With warmup, the mask rate starts low, rises steadily (to approximately 55% by epoch 10), and stabilizes at approximately 70%–75%, indicating improving pseudo-label quality. Without warmup, the mask rate is initially higher but saturates at a lower level, suggesting that early unreliable pseudo-labels limit final performance.

method avoids heatmap-based alignment, which requires storing intermediate spatial feature maps for each concept. In addition, **FixCBM** removes the dependency on architecture-specific spatial representations, making it simpler and more flexible to apply across different backbones. Overall, while incurring a modest increase in training time, **FixCBM** achieves substantially better memory efficiency and model simplicity, making it a practical alternative for semi-supervised concept learning.

5 Conclusion

We presented **FixCBM**, a semi-supervised framework for Concept Embedding Models that replaces spatial heatmap alignment with confidence-guided consistency learning in concept space. Our key insight is that the main challenge in semi-supervised CBMs lies in reliable concept-level supervision on unlabeled data rather than spatial alignment. By enforcing weak-to-strong consistency over concept predictions, **FixCBM** provides a simple and effective way to exploit unlabeled data without requiring intermediate spatial features or architecture-specific designs. This is achieved through three components: per-concept confidence masking, a concept-space consistency objective, and a warmup schedule that stabilizes early training. Experiments on CUB-200-2011, AwA2, PBC,

Table 4. Computational cost comparison on CUB-200-2011 (r=0.1, batch size 32). Time per epoch measured on a single NVIDIA RTX 4090.

Method	Time/Epoch (s)	GPU Mem (GB)	FLOPs/batch
CEM-SS (KNN only)	22.81	13.27	6.86
SSCBM [9]	39.78	14.29	13.71
FixCBM (Ours)	**46.3**	**5.13**	**13.74**

and 7-point show consistent improvements in both concept and task accuracy over prior baselines, while preserving strong intervention behavior. Moreover, **FixCBM** is more memory-efficient and avoids reliance on architecture-specific representations, making it a practical and flexible solution for semi-supervised concept learning.

References

1. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
2. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)
3. Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., Dvijotham, K.: Interactive concept bottleneck models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 6205–6212 (2023)
4. Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., Savvides, M.: Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. In: International Conference on Learning Representations (2023)
5. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703 (2020)
6. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shao, Z., Sourek, G., Stevenson, A., Tonda, A., et al.: Concept embedding models: Beyond the accuracy-explainability trade-off. In: Advances in Neural Information Processing Systems. vol. 35, pp. 21400–21413 (2022)
7. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI Magazine **38**, 50–57 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
9. Hu, L., Huang, T., Xie, H., Gong, X., Ren, C., Hu, Z., Yu, L., Ma, P., Wang, D.: Semi-supervised concept bottleneck models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2110–2119 (October 2025)
10. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE Journal of Biomedical and Health Informatics **23**, 538–546 (2018)

11. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International Conference on Machine Learning. pp. 5338–5348. PMLR (2020)
12. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representations (2017)
13. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: International Conference on Learning Representations (2023)
14. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: Advances in Neural Information Processing Systems. vol. 29 (2016)
15. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kumari, U., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: Advances in Neural Information Processing Systems. vol. 33, pp. 596–608 (2020)
16. Tan, A., Liang, J., Wu, W.Z., Zhang, J.: Semi-supervised partial multi-label classification via consistency learning. Pattern Recognition **131**, 108839 (2022)
17. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
18. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Transactions on Neural Networks and Learning Systems **32**, 4793–4813 (2020)
19. Tsutsui, S., Pang, W., Wen, B.: Wbcatt: A white blood cell dataset annotated with detailed morphological attributes. In: Advances in Neural Information Processing Systems. vol. 36, pp. 50796–50824 (2023)
20. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Caltech-ucsd birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology (2010)
21. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**, 2251–2265 (2018)
22. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6256–6268 (2020)
23. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: International Conference on Learning Representations (2023)
24. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Qin, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In: Advances in Neural Information Processing Systems. vol. 34, pp. 18408–18419 (2021)