Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36457

Cross-Resolution Knowledge Distillation for Low-Compute Fine-Grained Bird Classification

Chenqi Li

Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, China
t330034023@mail.bnbu.edu.cn

Abstract. Fine-grained visual classification relies on subtle local cues and is highly sensitive to input resolution, yet practical deployment often constrains image size and inference cost. Existing low-resolution recognition methods often depend on additional network structures or complex training modules, limiting deployment simplicity. This study evaluates a simple cross-resolution knowledge distillation (KD) strategy under a fixed low-compute constraint. The teacher receives higher-resolution inputs to provide richer class-discriminative visual knowledge, whereas the student is trained at 160×160 and is the only model retained for deployment. Both networks use ResNet-18, allowing resolution-aware supervision to be studied without changing the inference architecture. On the CUB-200-2011 dataset, the 160×160 baseline achieved 66.01±0.53% test accuracy. The 224→160 KD configuration improved accuracy to 70.15±0.63%, while the final 384→160 configuration with $T = 4$, $\alpha = 0.75$, and a 30-epoch cosine schedule reached 71.60±0.68% accuracy and 71.56±0.62% macro-F1. The results also show that a stronger teacher does not automatically produce a stronger student: transfer quality depends on the balance between teacher supervision strength and student optimization. These findings support cross-resolution KD as a practical way to recover fine-grained information during training while preserving low-resolution inference cost.

Keywords: fine-grained classification, knowledge distillation, low-resolution recognition, ResNet-18, CUB-200-2011

1. Introduction

Fine-grained image classification aims to distinguish visually similar subcategories, such as bird species, vehicle models, or aircraft variants. Unlike generic classification, the discriminative evidence is often concentrated in small regions, including feather color, beak shape, wing pattern, or other subtle part-level details. In mobile, embedded, remote-sensing, and bandwidth-constrained systems, however, images may need to be captured, transmitted, or processed at reduced resolution to satisfy memory and computation limits. This reduction can remove precisely the local visual cues that fine-grained recognition requires, creating a tension between recognition quality and efficient deployment. Convolutional backbones such as ResNet provide strong visual representations [1], but lowering the input resolution can substantially weaken the information available to a fixed-capacity classifier.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.36457

Knowledge distillation (KD) is a model-compression and knowledge-transfer framework in which a student model learns from both ground-truth labels and the predictive information produced by a teacher model [2]. The teacher is usually trained with greater model capacity, richer input information, or both, while the student is optimized to reproduce useful teacher behavior under a more restrictive deployment budget. Classical KD transfers softened class probabilities, and later work has extended the idea through contrastive representations [3], cross-stage feature review [4], decoupled target and non-target logits [5], relational predictions from stronger teachers [6], multi-level logit alignment [7], and standardized logit distributions [8]. These studies show that the form and strength of the transferred supervision can be as important as teacher accuracy itself.

Cross-resolution KD differs from conventional capacity-oriented distillation because teacher and student operate under different input-information conditions. A high-resolution teacher can observe finer local structures, while a low-resolution student must learn to exploit the corresponding class relationships from a compressed view. Guo et al. [9] treated image size as a dimension of model compression and transferred knowledge across different input resolutions. More recent work has targeted low-resolution fine-grained recognition directly: Liang et al. [10] introduced dynamic semantic structure distillation (DSSD), while Zhang et al. [11] combined augmented fine-grained samples with decoupled distillation. These methods demonstrate the value of resolution-aware supervision, but they also introduce additional structures or training components beyond a standard logit-based teacher-student objective.

This complexity motivates a narrower question: how much can a simple cross-resolution logit KD pipeline recover when the deployed architecture and low-resolution input are kept fixed? Isolating this setting helps evaluate whether resolution difference itself can act as a useful source of transferable knowledge without adding inference-time modules. The contributions of this study are threefold. First, it establishes a reproducible paired-resolution training protocol in which teacher and student receive spatially aligned views of the same image. Second, it evaluates standard logit-based KD under a fixed 160×160 ResNet-18 deployment constraint, separating training-time supervision from inference-time cost. Third, it systematically studies the effects of teacher input resolution, distillation temperature, teacher-loss weight, and training schedule on cross-resolution transfer.

2. Materials and methods

2.1. Dataset and preprocessing

The CUB-200-2011 dataset contains 11,788 images from 200 bird species [12]. It is well suited to this study because fine-grained bird recognition depends heavily on small local details, making resolution-induced information loss directly relevant to classification performance. The official split provides 5,994 training images and 5,794 test images. During development, the official training portion was divided into 5,094 training images and 900 validation images using a fixed stratified split with split seed 3407. The official test set was kept untouched until the final configuration had been selected, reducing test-set feedback during model selection. The three formal student-training seeds were 42, 123, and 2026. Bounding-box annotations and test-time augmentation were not used, so all results correspond to full-image classification under the same low-compute deployment setting.

Training pairs were generated from the same source image to preserve spatial correspondence between teacher and student views. A random resized crop was sampled once and shared by both branches, followed by the same horizontal flip and color perturbation. After the shared augmentation, the image was resized separately to the teacher and student resolutions. A shared

DOI: 10.54254/2755-2721/2026.36457

random-erasing region was also mapped proportionally across resolutions. Consequently, the 384×384 teacher and 160×160 student observed the same semantic crop rather than unrelated random views. Validation and test preprocessing were deterministic: each image was resized to approximately 1.14 times the target size, center-cropped to the required input resolution, converted to a tensor, and normalized using ImageNet statistics.

2.2. Cross-resolution knowledge distillation framework

The deployed classifier is an ImageNet-pretrained ResNet-18 with a 200-class output layer [1]. The student input is fixed at 160×160 for every experiment. Teacher models use the same ResNet-18 architecture but receive higher-resolution inputs of either 224×224 or 384×384. During distillation, the teacher is frozen, kept in evaluation mode, and executed without gradients; only the student is optimized. This design makes the resolution difference, rather than a larger teacher architecture, the principal source of additional training information. After training, the teacher is discarded, so no teacher computation, auxiliary branch, or feature-matching module is required at inference time.

For a teacher logit vector z^t and student logit vector z^s, temperature T produces softened class probabilities as follows:

$$p_i^t(T) = \frac{\exp(z_i^t/T)}{\sum_j \exp(z_j^t/T)}, \quad p_i^s(T) = \frac{\exp(z_i^s/T)}{\sum_j \exp(z_j^s/T)} \tag{1}$$

The distillation term is the Kullback-Leibler divergence between the softened distributions, scaled by T^2 :

$$L_{KD} = T^2 \cdot KL(p^t(T)||p^t(T)) \tag{2}$$

The final student objective combines label supervision and teacher supervision:

$$L = (1-\alpha)L_{CE} + \alpha L_{KD} \tag{3}$$

Here, L_{CE} is cross-entropy with label smoothing 0.1, T is the distillation temperature, and α controls how strongly the student follows the teacher distribution. A larger α assigns more weight to the teacher signal. Figure 1 summarizes the paired-resolution framework and shows that only the 160×160 student is retained after training.

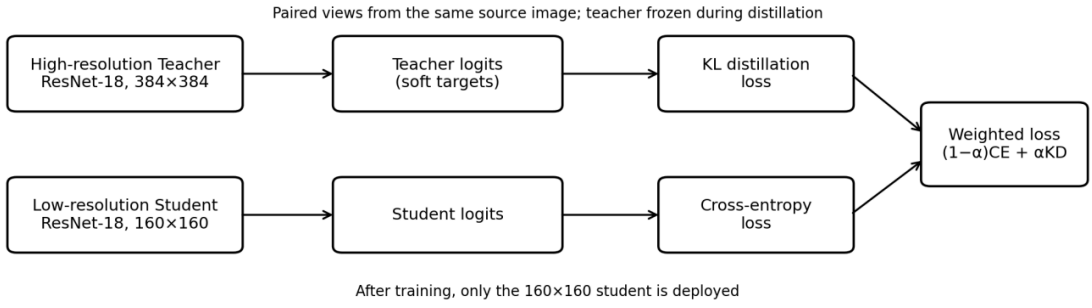

Figure 1. Final cross-resolution knowledge-distillation framework. The teacher is used only during training and deployment retains the 160×160 student (picture credit : original)

DOI: 10.54254/2755-2721/2026.36457

2.3. Experimental settings

The initial 224→160 protocol used $T = 4$ and $\alpha = 0.5$. Temperature was evaluated at $T \in \{2,4,8\}$, and α was evaluated at $\{0.25,0.50,0.75\}$ on seed 123. The later 384→160 experiments first retained $T = 4$ and $\alpha = 0.5$ to isolate the effect of using a higher-resolution teacher under the original objective. Because this change alone did not improve the three-seed mean, α was increased to 0.75. The final configuration used the 384×384 teacher, $T = 4$, $\alpha = 0.75$, and a 30-epoch cosine schedule.

All student models were trained with batch size 16 using AdamW, weight decay 10^{-4}, a classifier-head learning rate of 3×10^{-4}, and a backbone learning rate of 5×10^{-5}. The backbone was frozen for the first epoch and then unfrozen. The original protocol used a 20-epoch cosine schedule with minimum learning rate 10^{-6}; the final configuration extended the cosine horizon to 30 epochs. Early stopping used validation accuracy with patience 4. Model checkpoints were selected solely by validation accuracy. Accuracy and macro-F1 were reported, and three-seed results are summarized using the arithmetic mean and sample standard deviation.

2.4. Deployment constraint analysis

The deployment constraint is deliberately stricter than the training constraint. The 160×160 ResNet-18 baseline contains approximately 11.279 million parameters and requires about 0.925 GMAC per image in the project benchmark. Because the distilled student has exactly the same architecture and input resolution, knowledge distillation does not increase its parameter count or theoretical multiply-accumulate cost. Higher-resolution teacher computation is therefore a training-time expense only.

3. Results

3.1. Evaluation protocol and model selection

All methods were evaluated under the same class ordering and deterministic evaluation transform. During model development, only the 900-image validation split was used for checkpoint selection and hyperparameter decisions. The official 5,794-image CUB test split was evaluated only after the final 384→160 configuration had been locked. This protocol produced close agreement between development and final evaluation: the three-seed validation mean of the final method was 71.55±0.69%, while the official-test mean was 71.60±0.68%. The small difference indicates that the selected validation split did not create an obvious optimistic bias for the final configuration.

3.2. Effect of cross-resolution knowledge distillation

Table 1 summarizes the main official-test comparison. The low-resolution 160×160 ResNet-18 baseline achieved 66.01±0.53% accuracy. Distillation from the 224×224 teacher increased the mean to 70.15±0.63%, corresponding to an absolute improvement of 4.14 percentage points. The final 384→160 configuration reached 71.60±0.68%, improving the baseline by 5.59 percentage points and the original 224→160 KD system by 1.45 percentage points. For the final method, the official-test macro-F1 was 71.56±0.62%, closely matching accuracy and suggesting that the improvement was not driven only by a small subset of classes.

DOI: 10.54254/2755-2721/2026.36457

Table 1. Official CUB-200-2011 test accuracy across three training seeds. "pp" denotes percentage points

Method	Teacher input	Student input	Test Acc. (%)	Gain vs. baseline
160×160 baseline	—	160	66.01 ± 0.53	—
224→160 KD ($T = 4$, $\alpha = 0.5$)	224	160	70.15 ± 0.63	+4.14 pp
384→160 final KD	384	160	71.60 ± 0.68	+5.59 pp

The gain is relevant to the low-compute objective because the deployed network remains the same 160×160 ResNet-18 used by the baseline; only the training procedure changes. In contrast, increasing the deployment input from 160×160 to 224×224 would increase image area by approximately 1.96×, and moving to 384×384 would increase it by 5.76×. Cross-resolution KD therefore uses high-resolution information during training while avoiding the corresponding resolution cost at inference time.

3.3. Ablation

A higher-performing teacher was not sufficient by itself. The 384×384 teacher improved teacher validation accuracy by 5.00 percentage points over the 224×224 teacher (78.78% versus 73.78%). However, when the original $T = 4$, $\alpha = 0.5$ and 20-epoch protocol was retained, the 384→160 student achieved a three-seed validation mean of 70.00±0.78%, slightly below the 70.15±1.07% obtained by the original 224→160 system. This result motivates examining the interaction between teacher resolution and the strength of the distillation objective rather than treating teacher accuracy as the only controlling factor.

Table 2. Seed-123 ablation illustrating the interaction among teacher resolution, distillation weight, and schedule length

Teacher input	α	Max epochs	Seed	Val. Acc. (%)
224	0.50	20	123	69.00
224	0.75	20	123	69.33
384	0.50	20	123	69.67
384	0.75	20	123	71.11
384	0.75	30	123	71.33

The single-seed ablation in Table 2 explains the transition to the final configuration. With the 224×224 teacher, increasing α from 0.50 to 0.75 changed validation accuracy only from 69.00% to 69.33%. With the 384×384 teacher, the same increase in teacher weight changed the seed-123 result from 69.67% to 71.11%. Extending the cosine schedule from 20 to 30 maximum epochs then produced a smaller additional increase to 71.33%. In the earlier temperature study with the 224×224 teacher and $\alpha = 0.5$, $T = 4$ achieved 69.00% on seed 123, compared with 68.56% for $T = 2$ and 68.78% for $T = 8$, so $T = 4$ was retained.

DOI: 10.54254/2755-2721/2026.36457

3.4. Discussion

The ablation results suggest that cross-resolution transfer is controlled by more than teacher accuracy. A higher-resolution teacher contains richer discriminative information, but the low-resolution student must receive a sufficiently strong and learnable distillation signal to benefit from it. This interpretation is consistent with prior work that modifies logit relationships, target/non-target decomposition, or semantic structure to improve knowledge transfer [5-11]. The present study intentionally keeps the loss simple: its aim is not to claim a new state-of-the-art distillation algorithm, but to determine how much standard logit-based KD can recover under a fixed low-resolution deployment budget.

The results also support a broader interpretation: input-resolution difference itself can serve as a source of transferable knowledge. Because teacher and student share the same ResNet-18 architecture, the improvement cannot be explained by a larger deployed network. Instead, the teacher has access to finer visual evidence during training and communicates part of that information through its prediction distribution. This makes cross-resolution KD attractive when high-resolution computation is affordable offline but not at deployment time.

The comparison between the final 384→160 system and the original 224→160 system must nevertheless be interpreted carefully. Teacher resolution, α, and the training horizon all changed between these two final configurations. Therefore, the 1.45-point official-test advantage cannot be attributed to teacher resolution alone. A fully matched factorial experiment, for example, 224→160 and 384→160 both using $\alpha = 0.75$ and a 30-epoch schedule, would be required to isolate the causal contribution of resolution. In addition, only three student seeds and one 384×384 teacher seed were used. The results support a stable practical improvement, but they do not justify strong claims of statistical significance or universal superiority across datasets and backbones.

The absolute improvement should also be interpreted relative to the deliberately restricted student. Fine-grained recognition studies often maximize accuracy using larger input sizes, stronger backbones, part-localization modules, or specialized augmentation. Those alternatives were not the deployment target here. Under the fixed 160×160 ResNet-18 constraint, moving additional computation to training recovered useful class structure without introducing a second inference branch. The low standard deviations of 0.53%, 0.63%, and 0.68% for the baseline, 224→160 KD, and final 384→160 KD respectively further indicate that the ordering of the three systems was not driven by one unusually favorable run.

4. Conclusion

This study shows that simple cross-resolution knowledge distillation can improve fine-grained recognition while preserving a fixed low-compute deployment model. With the deployed network held at 160×160 ResNet-18, the final configuration reached 71.60±0.68% official-test accuracy, a 5.59-percentage-point gain over the low-resolution baseline, without increasing inference-time parameters or theoretical computation. More importantly, the results indicate that input-resolution difference itself can act as a transferable source of knowledge: a teacher can use finer visual evidence during training to guide a student that must operate on lower-resolution inputs.

The study remains limited to one dataset, one deployment backbone, three student seeds, and a single 384×384 teacher seed, and the final 224→160 and 384→160 systems were not controlled under a fully matched factorial design. Future work can therefore test matched teacher-resolution settings, additional fine-grained datasets, and stronger but deployment-neutral logit distillation methods. Within the present scope, the findings support cross-resolution KD as a simple and

DOI: 10.54254/2755-2721/2026.36457

practical strategy for recovering fine-grained classification performance while keeping low-resolution inference unchanged. More broadly, resolution can be viewed not only as a deployment constraint but also as a complementary source of training information, offering a useful direction for efficient recognition systems in computation-constrained environments.

References

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
[2] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv:1503.02531.
[3] Tian, Y., Krishnan, D., & Isola, P. (2020). Contrastive representation distillation. *International Conference on Learning Representations (ICLR)*.
[4] Chen, P., Liu, S., Zhao, H., & Jia, J. (2021). Distilling knowledge via knowledge review. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5008–5017.
[5] Zhao, B., Cui, Q., Song, R., Qiu, Y., & Liang, J. (2022). Decoupled knowledge distillation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11953–11962.
[6] Huang, T., You, S., Wang, F., Qian, C., & Xu, C. (2022). Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems, 35*.
[7] Jin, Y., Wang, J., & Lin, D. (2023). Multi-level logit distillation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24276–24285.
[8] Sun, S., Ren, W., Li, J., Wang, R., & Cao, X. (2024). Logit standardization in knowledge distillation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15731–15740.
[9] Guo, G., Zhang, D., Han, L., Liu, N., Cheng, M.-M., & Han, J. (2024). Pixel distillation: Cost-flexible distillation across image sizes and heterogeneous networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(12), 9536–9550. https://doi.org/10.1109/TPAMI.2024.3421277.
[10] Liang, M., Huang, S., & Liu, W. (2024). Dynamic semantic structure distillation for low-resolution fine-grained recognition. *Pattern Recognition, 148*, 110216. https://doi.org/10.1016/j.patcog.2023.110216.
[11] Zhang, H., Qiao, Y., & Wang, M. (2025). Data augmentation guided decouple knowledge distillation for low-resolution fine-grained image classification. In *Pattern Recognition and Computer Vision: PRCV 2024, Lecture Notes in Computer Science, 15034*, 379–392. Springer. https://doi.org/10.1007/978-981-97-8505-627.
[12] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. California Institute of Technology, Technical Report CNS-TR-2011-001.