NYCU CV2025 HW1

GitHub Link:
https://github.com/CEJiang/NYCU-Computer-Vision-2025-Spring-HW1

1. Introduction

In this assignment, we aim to build a robust image classification model for the dataset provided in HW1. To achieve strong generalization performance, we adopt a transfer learning approach using ResNeXt-101 pretrained on ImageNet, and fine-tune its fully connected layers on the target dataset.

To tackle challenges such as visual similarity between species, we introduce several techniques :

- **Strong data augmentations** to enhance model robustness under various lighting conditions, perspectives, and background clutter.
- **Label smoothing and progressive loss switching** to stabilize training during early epochs.
- **Focal Loss** to emphasize hard-to-classify examples and mitigate the effects of class imbalance.
- **Exponential Moving Average (EMA)** of model weights to improve validation stability and final performance.
- **Test-Time Augmentation (TTA)** to further boost prediction accuracy by aggregating results from multiple augmented views during inference.

Our pipeline includes class-balanced loss weighting, cosine learning rate scheduling, and detailed monitoring through training curves and confusion matrices.

The final model achieves over **92–95% validation accuracy**, with **smooth convergence and minimal overfitting**.

2. Method

2.1 Data Preprocessing

RandomResizedCrop	Size = 224
RandomAffine	Degree = 30 Translate = [0.15, 0.15] Scale = [0.7, 1.2]
RandomHorizontalFlip	p = 0.5
ColorJitter	Brightness = 0.25 Contrast = 0.25 Saturation = 0.25 Hue = 0.125
RamdomGrayscale	p = 0.1
RandomApply(GaussianBlur)	p = 0.3
ToTensor	
RandomErasing	p = 0.2
Normalize	Mean = [0.485, 0.456, 0.406] Std = [0.229, 0.224, 0.225]

1. RandomResizedCrop

Randomly crops the image to a target size (e.g., 224×224) after rescaling it with a random scale factor. This helps simulate different object scales and compositions.

2. RandomAffine

Applies random affine transformations including rotation, translation, and scaling, which improves robustness to viewpoint and positional variations.

3. RandomHorizontalFlip

Randomly flips the image horizontally with a given probability. This is useful for classes that are symmetric or appear in either orientation.

4. ColorJitter & RandomGrayscale

- ColorJitter: Randomly changes the brightness, contrast, saturation, and hue of the image to simulate different lighting conditions.

- RandomGrayscale: Converts the image to grayscale with a given probability, helping the model learn to recognize shapes and textures without relying on color.

5. RandomApply(GaussianBlur)

Applies a Gaussian blur with random strength and kernel size (sigma=(0.1, 1.0)) to simulate out-of-focus or low-quality images. This operation is applied with 30% probability, helping the model learn to be robust to blurry or distorted inputs.

6. ToTensor

Converts the image from PIL format (used by ImageFolder) to a PyTorch tensor. This changes the shape to [C, H, W] and scales pixel values from [0, 255] to [0.0, 1.0], which is required by most PyTorch models including ResNeXt.

7. RandomErasing

Randomly erases a rectangular region of the image. This forces the model to learn from incomplete visual information and improves robustness against occlusion.

8. Normalize

All inputs were normalized with the ImageNet mean and std for pretrained compatibility.*

2.2 Model Architecture

Backbone: ResNext-101

Pre-trained Weights: ResNeXt101_32X8D_Weights.IMAGENET1K_V2

Model Fully connected Layer:

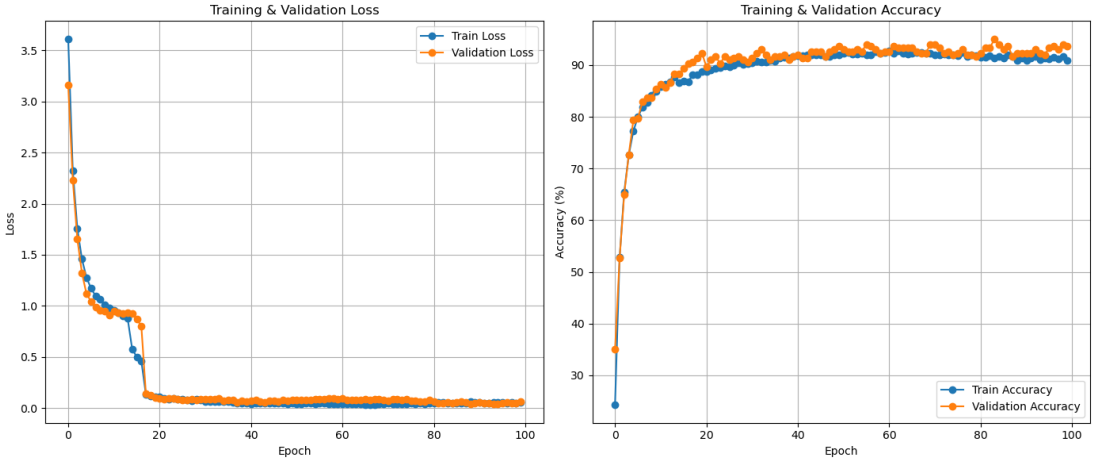
```
net.fc = nn.Sequential(
    nn.Linear(net.fc.in_features, 2048),
    nn.BatchNorm1d(2048),
    nn.ReLU(),
    nn.Dropout(0.4),


    nn.Linear(2048, 1024),
    nn.BatchNorm1d(1024),
    nn.ReLU(),
    nn.Dropout(0.3),

    nn.Linear(1024, 1024),
    nn.BatchNorm1d(1024),
    nn.ReLU(),
    nn.Dropout(0.2),

    nn.Linear(1024, num_classes)
)
```

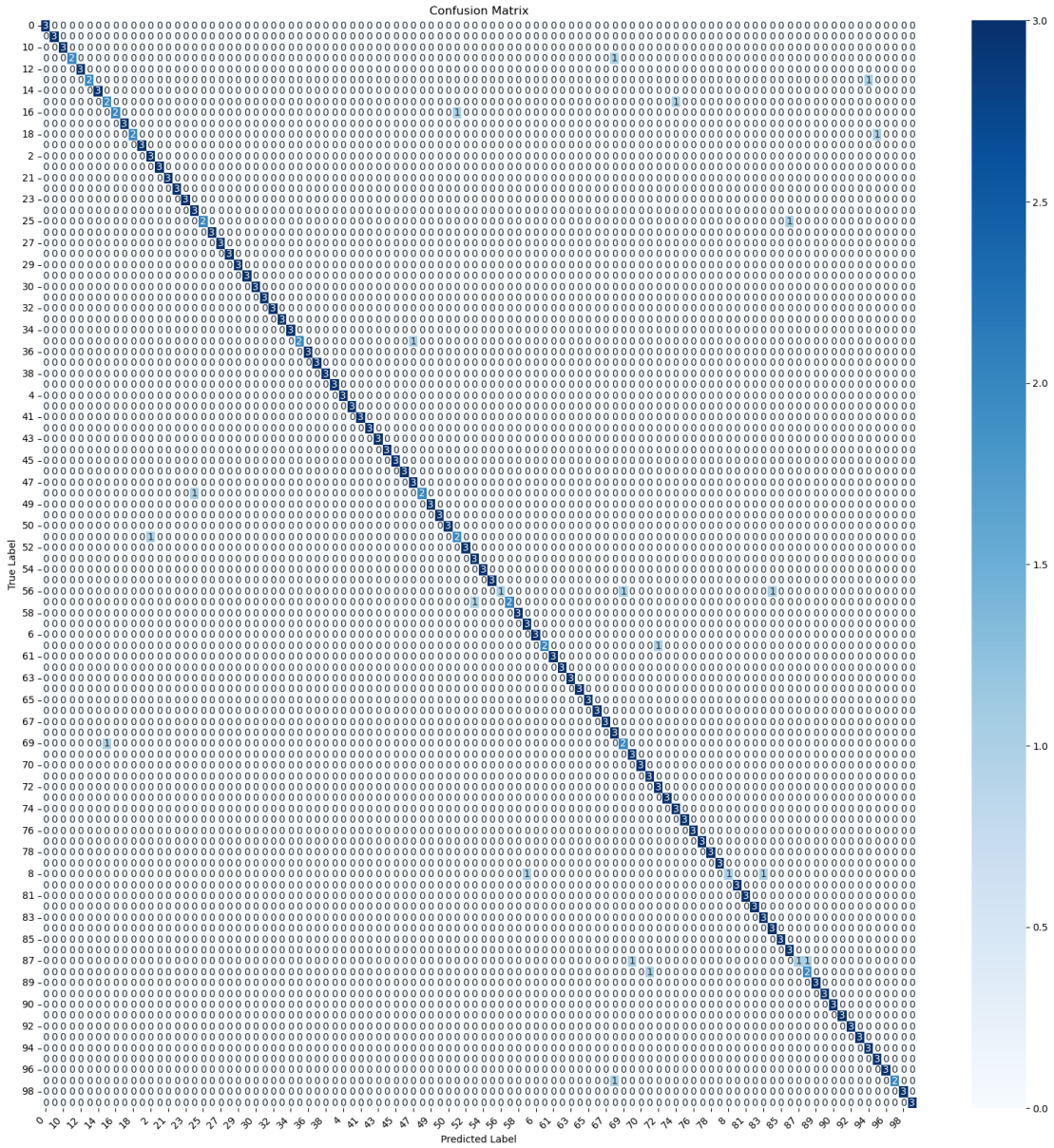
Total Parameters: 94.20M

2.3 Hyperparameters

Model	ResNeXt-101
Pretrained Weight	IMAGENET1K_V2
Learning rate	5×10^{-5}
Batch size	64
Epoch	100
Optimizer	AdamW
Eta_min	1×10^{-5}
T_max	50
Scheduler	CosineAnnealing
Label_smoothing	0.05
Criterion	CrossEntropy → SmoothFocal → Focal

3. Results


Training & Validation Loss
Train Loss
Validation Loss
Loss
Epoch
Training & Validation Accuracy
Train Accuracy
Validation Accuracy
Accuracy (%)
Epoch



The model achieves high validation accuracy (over 92%), with the best accuracy reaching 95%, and demonstrates very stable and smooth convergence as shown in the training curve. Throughout the training process, validation accuracy closely follows training accuracy with minimal gap or fluctuation in later epochs. The validation loss also remains low and consistent, indicating strong generalization ability.

The confusion matrix shows that predictions are highly concentrated along the diagonal, indicating very few misclassifications. This demonstrates the model's robustness and its ability to accurately distinguish between visually similar classes.

4. References

No references were used in this assignment.

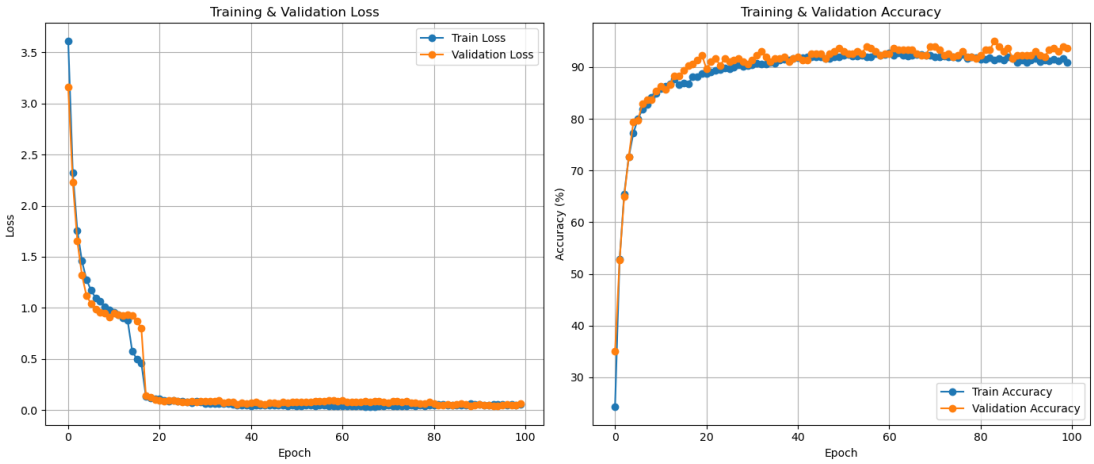
5. Additional experiments

	☑ Focal Loss	☒ Focal Loss
☑ EMA	Fig 5-1.	Fig 5-3.
☒ EMA	Fig 5-2.	Fig 5-4.

Hypothesis 1: Some classes may be harder to distinguish due to visual similarity. Focal Loss helps the model focus more on these challenging examples, potentially improving overall generalization.

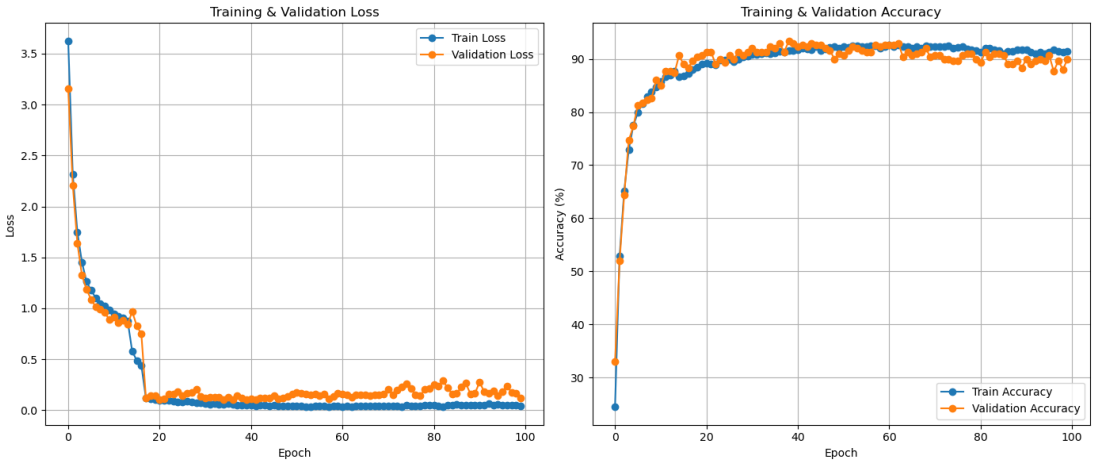
Hypothesis 2: EMA can stabilize the model weights and reduce variance during evaluation, especially at the later stages of training, leading to smoother and more reliable validation performance.

Hypothesis 3: Combining both techniques will allow the model to learn more robustly: Focal Loss for sample weighting and EMA for evaluation stability.



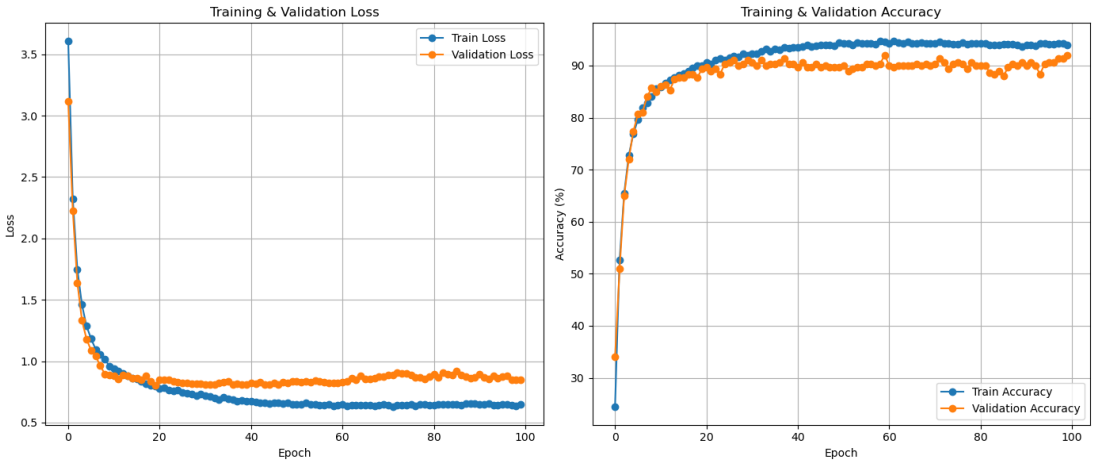
☑ EMA ☑ Focal Loss

Fig 5-1.



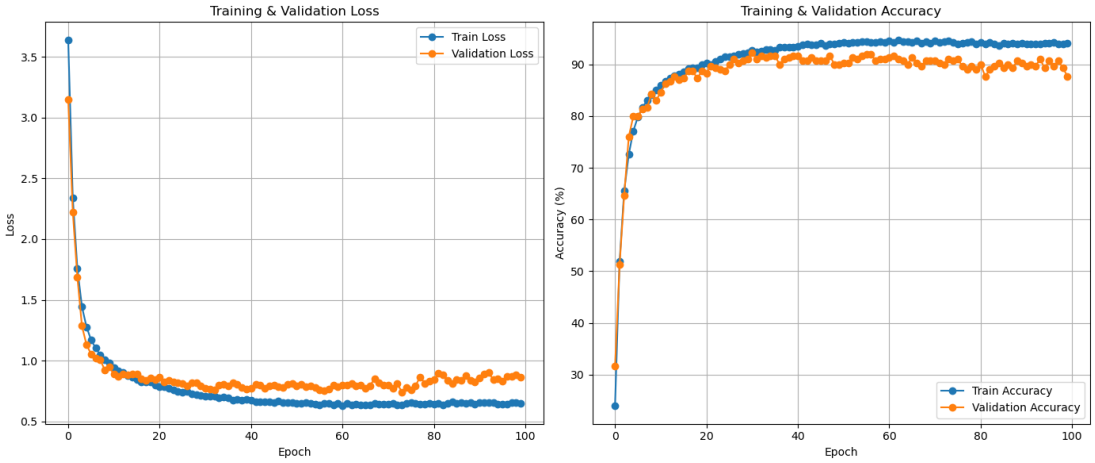
☒ EMA ☑ Focal Loss

Fig 5-2.



☑ EMA ☒ Focal Loss

Fig 5-3.



☒ EMA ☒ Focal Loss

Fig 5-4.

Hypothesis 1

Some classes may be harder to distinguish due to visual similarity. Focal Loss helps the model focus more on these challenging examples, potentially improving overall generalization.

How this may (or may not) work

Focal Loss down-weights easy examples and focuses the learning on harder cases by modifying the standard cross-entropy loss with a factor that emphasizes uncertain predictions. This can help the model better distinguish classes that are visually similar.

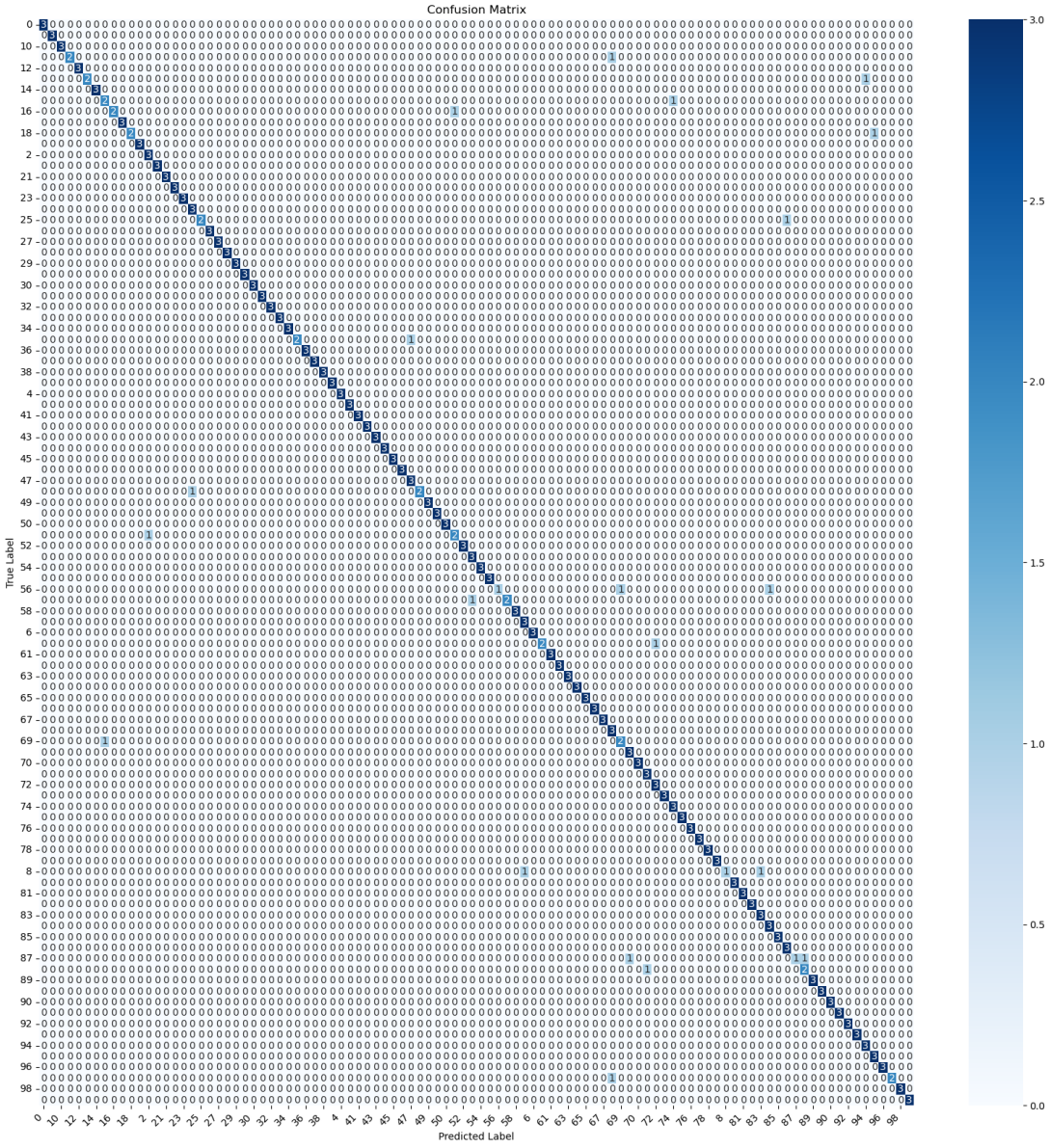
However, if most samples are already classified correctly early in training, or if the class distribution is balanced, the effect of Focal Loss might be minimal or even introduce instability by overemphasizing outliers.

Results and their implications

In Figure 5-1, where Focal Loss is applied, we observe a slightly improved validation accuracy and notably more stable validation loss compared to Figure 5-3, which uses CrossEntropyLoss. The training curve also shows that the model with Focal Loss converges faster and maintains smoother loss behavior throughout training.

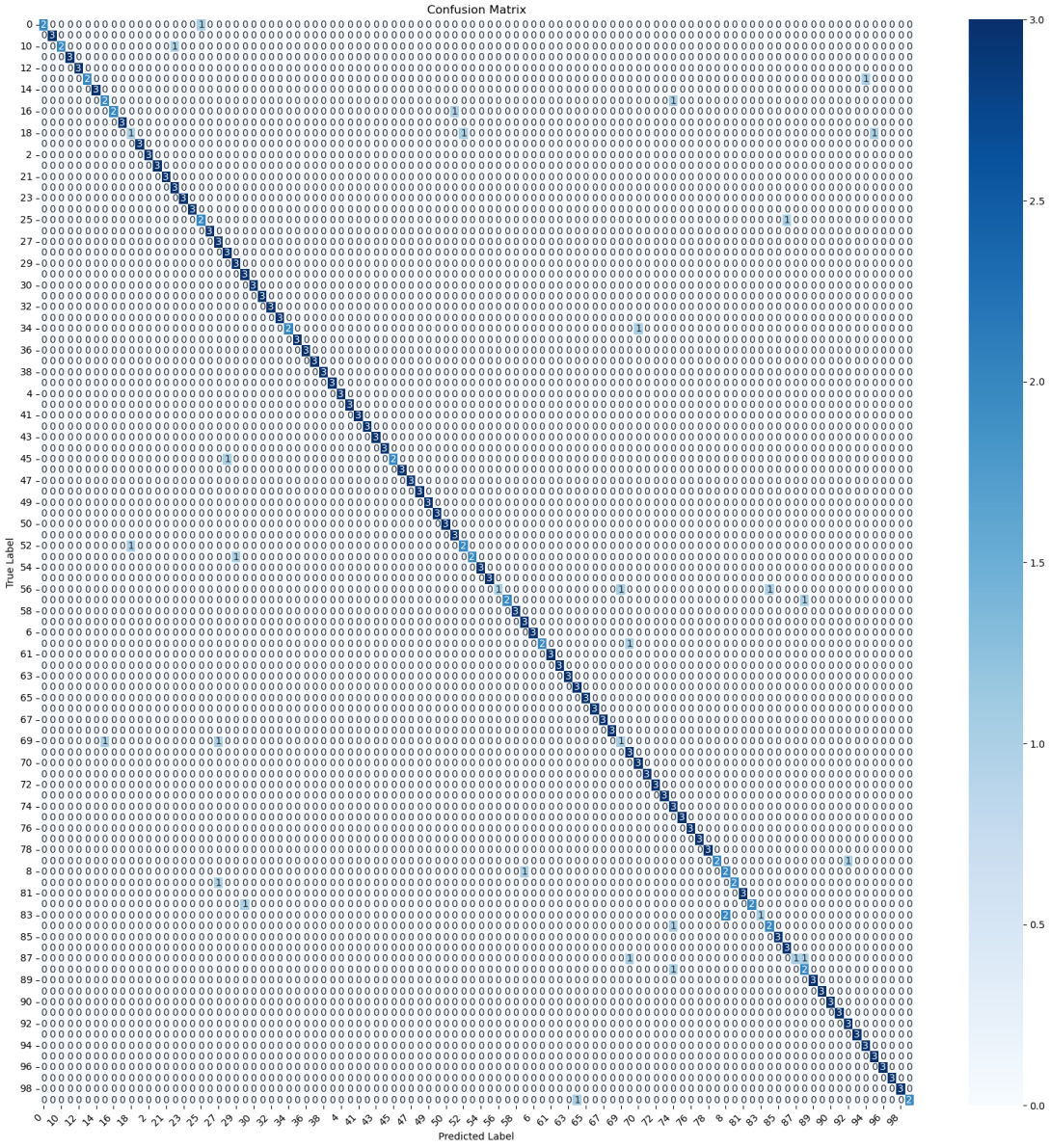
The confusion matrix in Figure 5-5.1 (with Focal Loss) also demonstrates slightly fewer misclassifications, particularly in off-diagonal elements, compared to Figure 5-5.2 (without Focal Loss). This suggests that Focal Loss enables the model to better handle hard-to-classify or visually ambiguous categories.

Although the overall accuracy improvement is marginal (around 0.5–1%), the use of Focal Loss leads to stronger generalization, fewer incorrect predictions, and more consistent performance across challenging samples. **These findings support the hypothesis 1 that Focal Loss helps the model focus on difficult samples, effectively reducing overfitting and enhancing robustness.**



☑ EMA ☑ Focal Loss

Fig 5-5.1.



☑ EMA ☒ Focal Loss

Fig 5-5.2.

Hypothesis 2

EMA can stabilize the model weights and reduce variance during evaluation, especially at the later stages of training, leading to smoother and more reliable validation performance.

How this may (or may not) work

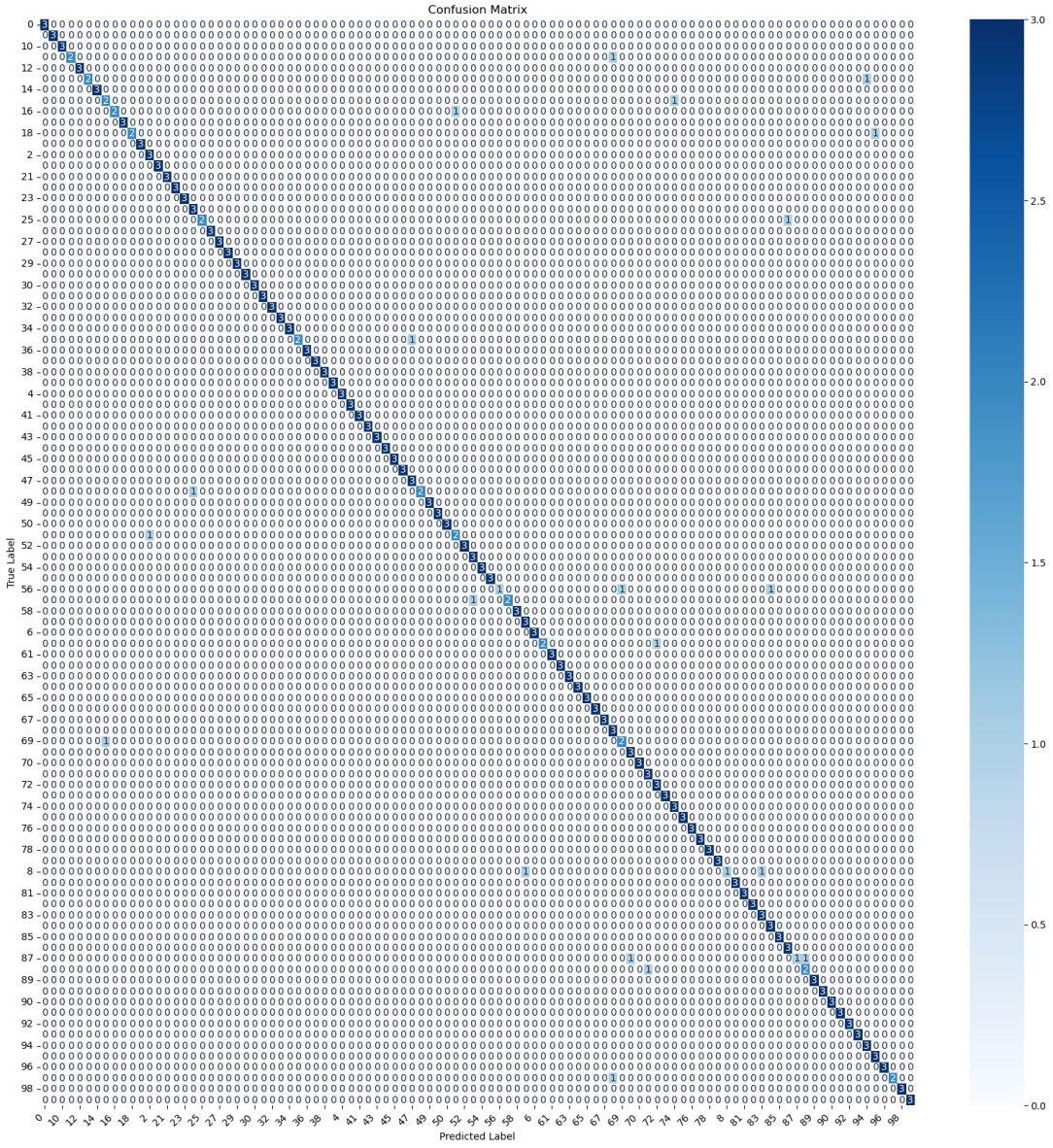
EMA works by maintaining a moving average of model parameters throughout training. At inference time, instead of using the latest (possibly noisy) parameters, the EMA-smoothed weights are used, which helps suppress sudden fluctuations in model performance.

However, if the model has already converged smoothly or if the training data is very clean, the benefits of EMA may be marginal or negligible.

Results and their implications

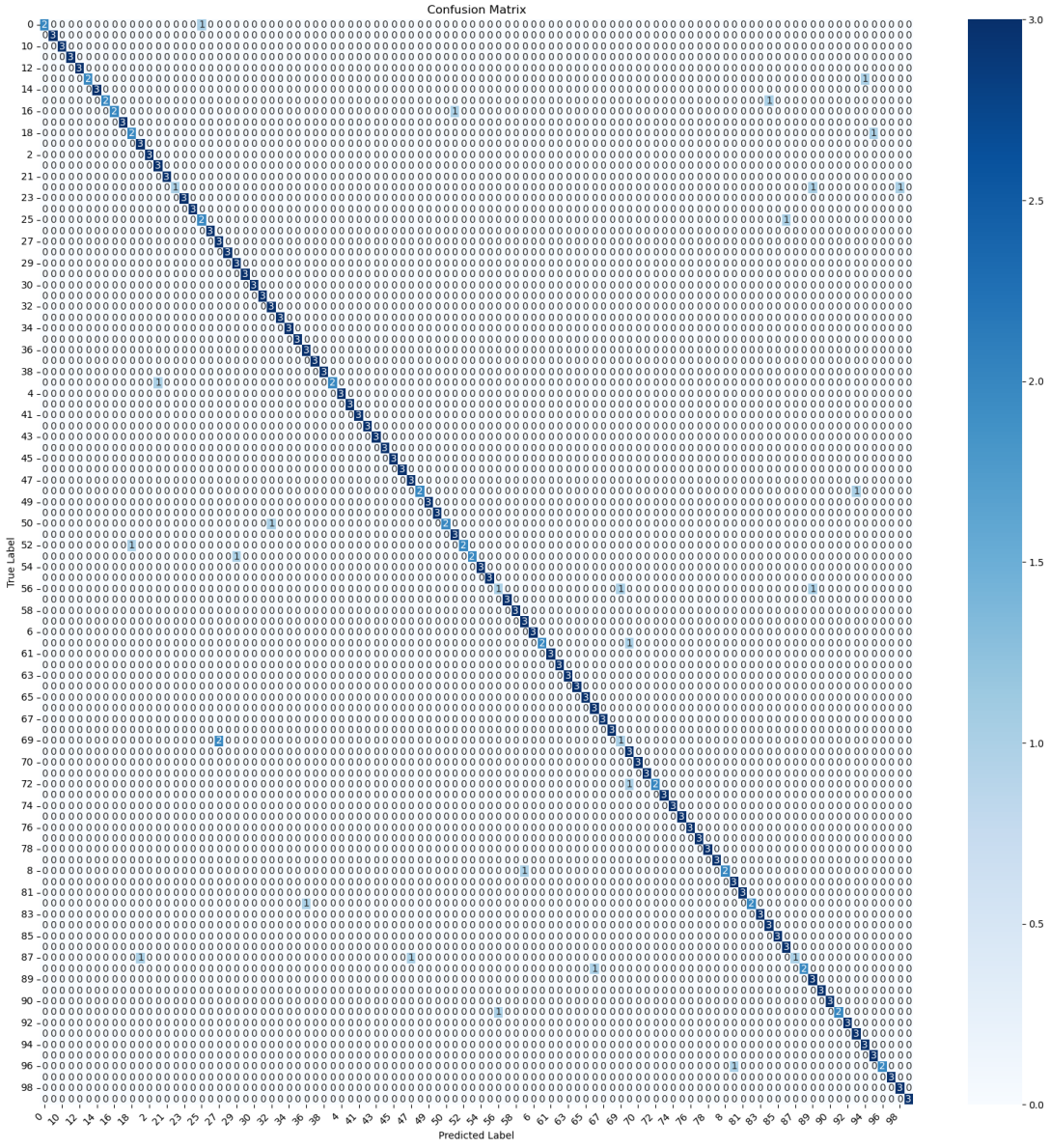
As shown in Figure 5-1 (with EMA) and Figure 5-2 (without EMA), both models achieve similar accuracy, but the EMA model demonstrates more stable validation loss curves, especially after 60 epochs. This implies the model generalizes better and avoids noisy predictions.

The confusion matrices in Figure 5-6.1 (with EMA) and Figure 5-6.2 (without EMA) also support this—while both models perform well, the EMA version has slightly fewer off-diagonal errors, showing better consistency in predictions, especially for borderline or confusing classes.



☑ EMA ☑ Focal Loss

Fig 5-6.1.



☒ EMA ☑ Focal Loss

Fig 5-6.2.

Hypothesis 3

Combining both techniques will allow the model to learn more robustly: Focal Loss for sample weighting and EMA for evaluation stability.

How this may (or may not) work

Focal Loss enhances learning from difficult samples, which often contribute disproportionately to validation errors. EMA reduces evaluation noise by smoothing parameter updates across epochs. When used together, these two techniques complement each other: Focal Loss focuses on what to learn (challenging samples), while EMA improves how the learned weights behave during evaluation (stability).

However, since both methods introduce regularization in different forms, combining them may lead to over-regularization, which could hinder learning if the model already converges well. In such cases, the performance gain might be marginal.

Results and their implications

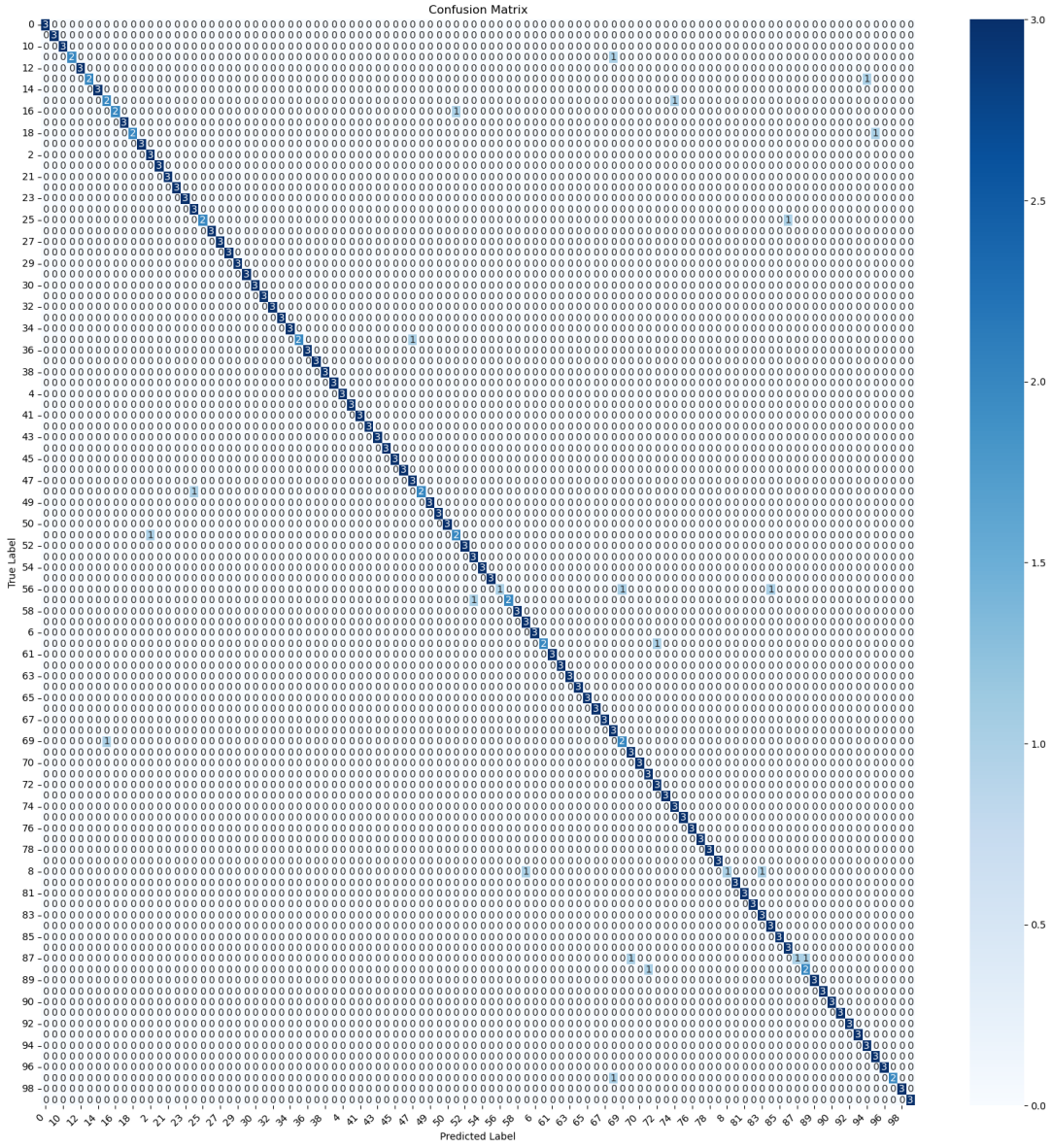
In Figure 5-1 (using both EMA and Focal Loss), we observe that the model converges quickly in the early stages and maintains stable validation accuracy throughout training. The validation loss is also smoother and lower compared to Figure 5-4 (without EMA and Focal Loss).

Looking at the confusion matrices, Figure 5-7.1 (with EMA and Focal Loss) shows predictions that are more concentrated along the diagonal, indicating fewer misclassifications. In contrast, Figure 5-7.2 (baseline) exhibits more errors, especially among visually similar categories.

Overall, the combined model:

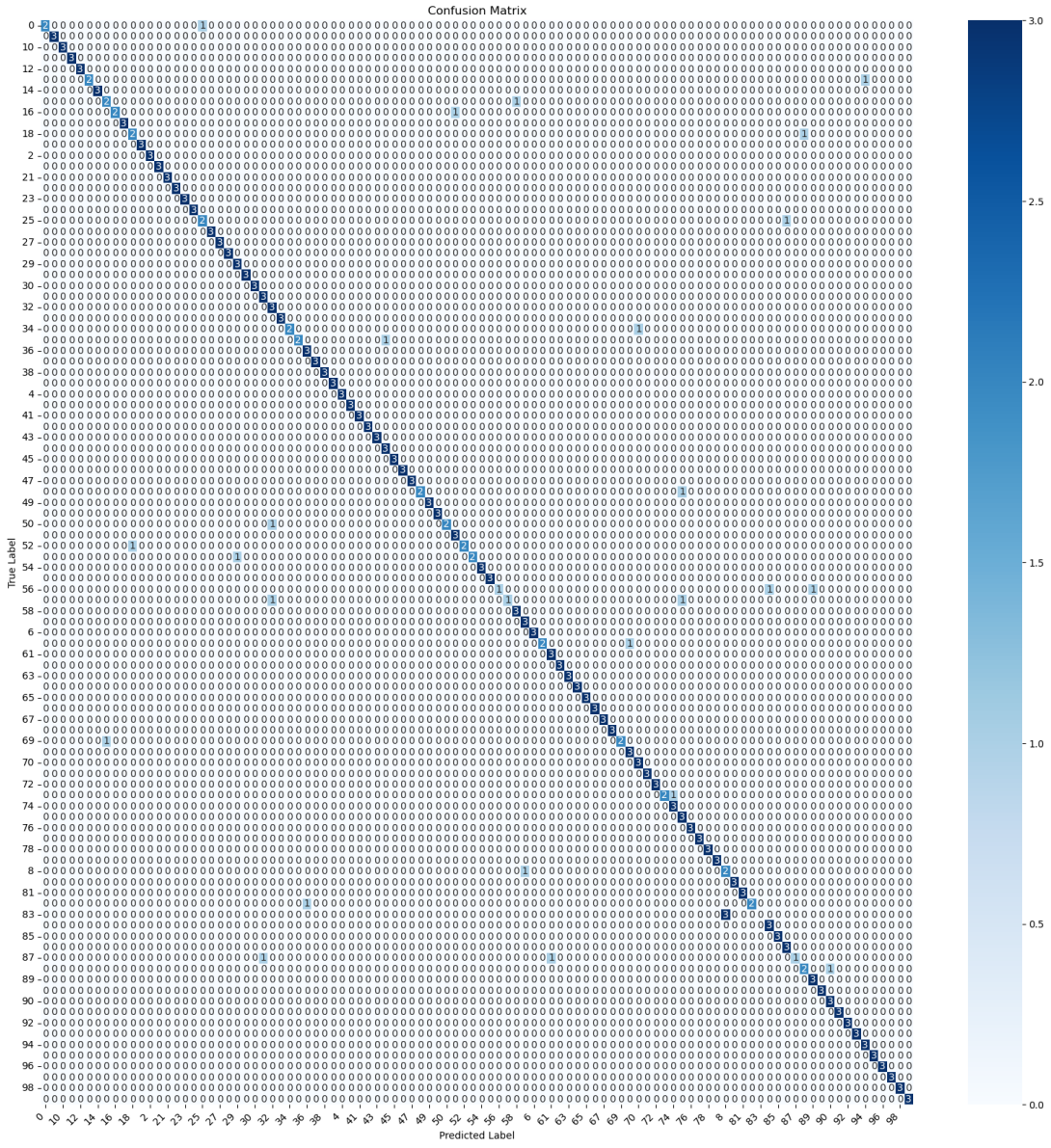
- Converges rapidly in the early stage
- Shows smoother and more stable validation performance in later epochs
- Makes fewer mistakes on hard-to-classify samples, as seen in the confusion matrix

Although the overall validation accuracy gain is modest (around 1–2%), the lower variance and reduced misclassifications suggest stronger generalization and improved robustness, especially for ambiguous visual classes.



☑ EMA ☑ Focal Loss

Fig 5-7.1.



☒ EMA ☒ Focal Loss

Fig 5-7.2.