Learning from Synthetic Data via Provenance-Based Input Gradient Guidance

Supplementary Material

Table 8. Hyperparameters of each dataset during training.

Training dataset	UCF101-24 [37]	CUB [42]			iWildCam [20]	Waterbirds [31]
Backbone	HRNet	VGG16	DeiT-S	ResNet-50	ResNet-50	ResNet-50
Mixing probability	1.0	1.0	0.79	-	-	-
Amount of Augmented Data Added.	-	-	-	1000	2224	839
Loss weight α in Eq. (1)	0.01	0.05	0.01	0.1	0.1	0.1
Optimizer	SGD	SGD	AdamW	SGD	SGD	SGD
Number of epochs	40	15	10	20	10	20
Batch size	30	32	32	128	128	128
Learning rate	7.5e-3	1e-2	1e-5	1e-3	1e-3	1e-3
LR scheduler	linear	linear	linear	cosine	cosine	cosine
Weight decay	2.5e-5	5e-4	5e-3	1e-5	1e-4	1e-4
Momentum	0.9					

6. Implementation Details

In this section, we provide details on data augmentation and training hyperparameters.

As described in Sec. 3.2.3, provenance information is derived by computing a difference image between the generated and source images, followed by Otsu binarization [22], to produce a binary mask distinguishing target regions from non-target regions.

As illustrated in Fig. 6, paired examples of synthetic-data samples alongside their corresponding provenance masks are visualized. Only a subset of these difference-based masks accurately captures the true target regions; many include residual background or synthetic artifacts. Improving provenance information quality, for example by leveraging cross-attention signals from the image generation model instead of relying solely on difference images, is an important direction for future work.

6.1. Hyperparameters

The hyperparameters for each dataset and synthesis setting are summarized in Tab. 8. Two types of synthesis are considered: mixing-based synthetic learning methods for localization and image editing methods for classification.

For the mixing-based synthetic learning methods (BatchMix and CutMix in Secs. 4.3.1 and 4.3.2), skeleton inputs with an HRNet backbone are used for weakly supervised spatio-temporal action localization on UCF101-24, and image inputs with VGG16 and DeiT-S backbones are used for weakly supervised object localization on CUB. The mixing probability (second row of Tab. 8) is set to 1.0 for UCF101-24, and to 1.0 and 0.79 for CUB with VGG16 and DeiT-S backbones, respectively.

For image editing methods (ALIA in Sec. 4.3.3) on CUB, iWildCam, and Waterbirds, we use ResNet-50 as the backbone for image classification. Instead of using a mixing probability, we control the number of generated images added to each training set ("Amount of Augmented Data Added" in Tab. 8). Following ALIA [10], we generate 1000 additional images for CUB, 2224 for iWildCam, and 839 for Waterbirds.

The loss balancing weight α in Eq. (1) is selected for each dataset and task (see Tab. 8) and kept fixed during training. For skeleton-based localization on UCF101-24, we train HRNet with SGD for 40 epochs using a linear schedule. For weakly supervised object localization on CUB, we train VGG16 for 15 epochs with SGD and DeiT-S for 10 epochs with AdamW, both using a linear schedule. For experiments with the image editing method on CUB, iWildCam, and Waterbirds, we train ResNet-50 with SGD for 20, 10, and 20 epochs, respectively, using a cosine schedule. The batch size, learning rate, weight decay, and optimizer are provided in Tab. 8.

Across all settings, we use a momentum of 0.9. The learning rate, weight decay, and α are tuned via coarse-to-fine grid search, while other hyperparameters follow standard practice for each backbone and dataset.

7. Ablation Study

7.1. Training Efficiency

As accuracy results are presented in Secs. 4.3.1 and 4.3.3, this section focuses on training efficiency. We measure efficiency by the number of epochs required to reach peak val-

idation performance ("Best epoch") under identical setups in Sec. 4.2, as summarized in Tabs. 9 and 10.

7.1.1. Image mixing

Compared with CutMix, our method reduces the Best epoch from $50 \rightarrow 15$ on VGG16 ($\approx 3.3\times$ fewer epochs) and from $30 \rightarrow 10$ on DeiT-S ($3\times$ fewer). This consistent reduction indicates faster and more stable optimization across both CNN and transformer backbones.

7.1.2. Image Editing by Image Generation Models

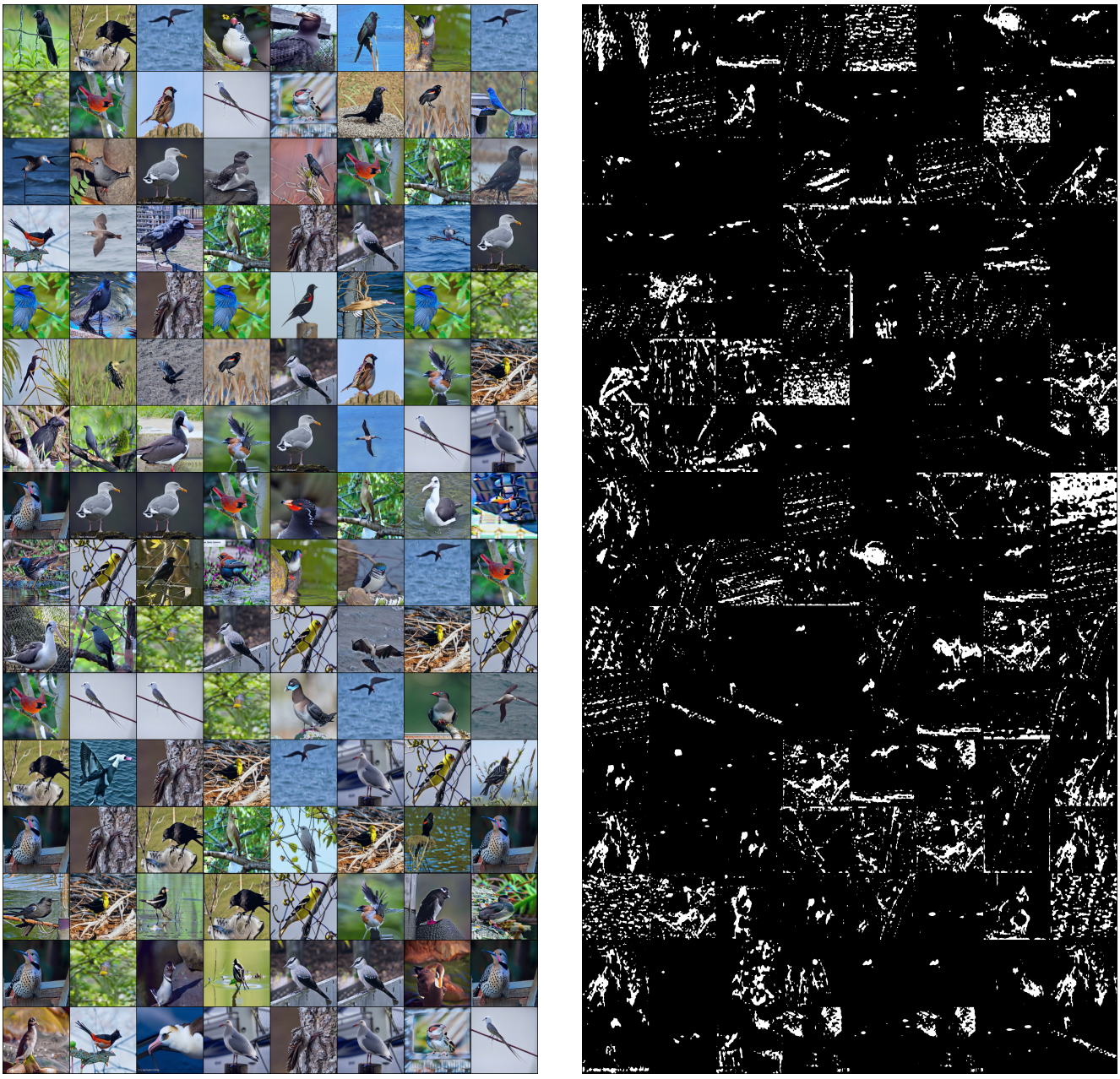
On CUB, our method reaches peak performance in 10 epochs, compared with 15 for ALIA ($1.5\times$ faster). On iWildCam, it converges in 5 epochs, compared with 10 for ALIA ($2\times$ faster). On Waterbirds, it reaches peak performance in 10 epochs, compared with 15 for ALIA ($1.5\times$ faster). These consistent trends across datasets with different distribution shifts suggest that provenance-guided regularization improves sample efficiency.

Table 9. Accuracy and training efficiency comparison of mix-based image synthesis (WSOL on CUB).

Method	Backbone	Acc. (%) ↑	Best epoch ↓
CutMix	VGG16 [35]	62.3	50
+Ours		65.1	**15**
CutMix	DeiT-S [40]	91.5	30
+Ours		92.0	**10**

Table 10. Accuracy and training efficiency comparison on CUB, iWildCam, and Waterbirds for the image editing method.

Method	CUB		iWildCam		Waterbirds	
	Acc. (%) ↑	Best epoch ↓	Acc. (%) ↑	Best epoch ↓	Acc. (%) ↑	Best epoch ↓
ALIA	71.7	15	77.7	10	71.4	15
+Ours	72.0	**10**	80.8	**5**	80.7	**10**



(a) Generated images from image editing synthesis.

(b) Provenance masks derived from difference images.

Figure 6. Visualization of image editing synthesis and the corresponding provenance masks.