Error-Aware Baselines for Generative Visual Recognition: A CIFAR-10 Study from Handcrafted Features to Transfer Learning

Anonymous CVPR submission

Paper ID *****

Abstract

Generative image models are increasingly used as data sources for visual recognition, but their value is difficult to judge without strong real-data baselines and careful error analysis. This paper presents an empirical CIFAR-10 study designed as a baseline and diagnostic framework for generative visual recognition. Starting from an existing classification report, we reorganize the study into a double-blind CVPR workshop style paper and evaluate a broad spectrum of methods: HOG features with KNN, SVM, and random forests; convolutional neural networks including VGG19-BN, ResNet, Inception-v4, and DenseNet; Pyramid Vision Transformer; deep-feature extraction followed by classical classifiers; and ImageNet-based transfer learning. The results show a consistent transition from handcrafted descriptors to learned representations: HOG-based methods reach at most 55.13% test accuracy, optimized ResNet reaches 93.60%, and DenseNet transfer learning reaches 93.67%. However, high average accuracy does not remove systematic failure modes. Confusion matrices and high-confidence errors reveal persistent ambiguities between cat/dog, deer/horse, automobile/truck, and airplane/ship, often caused by low resolution, background correlation, and semantic overlap. We use these findings to define an error-aware protocol for future generative augmentation: synthetic images should be evaluated not only by overall accuracy gains, but also by their ability to reduce class-wise confusion without introducing label or background bias.

1. Introduction

Image classification remains one of the clearest ways to study visual representation learning. Even though modern networks can obtain strong performance on small benchmarks, datasets such as CIFAR-10 [12] are still useful because they expose the interaction between feature design, model capacity, optimization, transfer learning, and class ambiguity. CIFAR-10 contains only 32×32 images, but its ten categories include several visually overlapping pairs. The low resolution makes the benchmark a compact test bed for asking whether a representation captures object semantics or relies on coarse texture and background cues.

At the same time, generative image models have rapidly changed the way visual data can be produced and reused. Generative adversarial networks [6], variational autoencoders [11], and diffusion models [9, 13] have demonstrated increasingly strong image synthesis capabilities. This progress has motivated a natural question: can synthetic images generated by modern generative models improve downstream visual recognition? Prior work has explored synthetic data for representation learning, robustness, and data augmentation [1, 8, 14], but its benefit is often task-dependent. Synthetic images may improve diversity and reduce overfitting, but they may also introduce distribution shift, label noise, background bias, or unrealistic class semantics.

In this work, we revisit CIFAR-10 as a controlled small-scale benchmark for studying this question from the recognition side. Rather than beginning with a generative model and assuming that additional images should help, we first build a set of classical and deep classification baselines to understand where errors occur and which types of visual ambiguity remain difficult. Our study begins with handcrafted HOG features [4] combined with classical classifiers, including K-nearest neighbors, support vector machines [3], and random forests [2]. We then compare several neural architectures, including VGG [15], ResNet [7], Inception [16, 17], DenseNet [10], and Pyramid Vision Transformer [18]. We also examine deep features with non-neural classifiers and transfer learning from ImageNet [5]. Finally, we analyze error cases and formulate a generative augmentation protocol for evaluating the effect of class-conditioned synthetic images.

Our motivation is that small-scale classification provides a simple but informative setting for generative augmentation. If synthetic images are useful, they should not only improve overall accuracy but also reduce meaningful class confusions. For example, a useful augmentation method should help

distinguish cats from dogs, deer from horses, automobiles from trucks, and airplanes from ships, rather than merely increasing the number of training images. Conversely, if synthetic images emphasize spurious background cues or produce semantically inconsistent samples, they may worsen exactly these ambiguous cases. This distinction is central for generative visual recognition: the value of generated images should be judged by their effect on the decision boundary, not only by visual realism or sample count.

This paper makes three main contributions:

- We provide a unified empirical comparison of handcrafted features, classical classifiers, deep networks, transformer-based models, deep-feature classifiers, and transfer learning for CIFAR-10 classification.
- We turn the original project results into a compact error analysis, using confusion matrices and high-confidence failure cases to identify persistent ambiguities relevant to synthetic data evaluation.
- We propose an error-aware generative augmentation protocol for studying when class-conditioned synthetic images help small-scale visual recognition and when they may introduce semantic or background bias.

The goal of this workshop paper is not to claim that synthetic data always improves CIFAR-10 classification. The available experiments are real-data recognition baselines and diagnostic analyses. We therefore position the generative component as a protocol and set of evaluation criteria for the next experimental stage. This is a useful contribution for a non-archival workshop setting because it connects an existing recognition study to the central question of how generative models can benefit computer vision, while keeping the claims faithful to the evidence.

2. Related Work

2.1. Classical Image Features and Classifiers

Before the success of deep learning, visual recognition systems often relied on handcrafted descriptors and classical machine learning models. HOG [4] represents local gradient orientation statistics and has been widely used for object detection and recognition. Classical classifiers such as K-nearest neighbors, support vector machines [3], and random forests [2] provide simple but useful baselines for evaluating the discriminative power of image features. Although these methods are generally outperformed by modern deep networks, they remain valuable for understanding the transition from handcrafted to learned representations.

2.2. Deep Networks for Image Classification

Convolutional neural networks have substantially improved image classification performance. VGG [15] demonstrated the effectiveness of using deep networks with small convolutional filters. ResNet [7] introduced residual connections, enabling much deeper models to be optimized effectively. Inception networks [16, 17] explored multi-branch convolutional modules and factorized convolutions to improve computational efficiency. DenseNet [10] further encouraged feature reuse by connecting each layer to all subsequent layers. More recently, transformer-based models have been adapted to vision tasks. Pyramid Vision Transformer [18] introduces a hierarchical transformer backbone and spatial-reduction attention, making transformer representations more suitable for dense prediction and visual recognition.

2.3. Transfer Learning

Transfer learning has become a standard technique for improving performance when the target dataset is limited. Models pretrained on large-scale datasets such as ImageNet [5] often learn transferable visual representations that can be adapted to downstream tasks. A common strategy is to initialize a network with pretrained weights and fine-tune either the classifier head or the full network. This approach has been widely used in image classification and visual recognition [19, 20]. In our study, transfer learning serves as a strong baseline for evaluating whether additional synthetic images are useful beyond pretrained representations.

2.4. Generative Models and Synthetic Data

Generative models have achieved significant progress in image synthesis. GANs [6] introduced adversarial learning for generating realistic images. VAEs [11] provided a probabilistic latent-variable framework for generative modeling. Diffusion models [9] and latent diffusion models [13] have further improved image quality and controllability. These advances have motivated the use of synthetic images for training visual recognition models. Recent work has investigated whether generated images can replace or complement real images for downstream classification and representation learning [1, 8, 14]. However, the effectiveness of synthetic data depends on image fidelity, diversity, label consistency, and distribution alignment with the target task. Our work focuses on this issue in a controlled CIFAR-10 setting.

3. Method

3.1. Problem Setup

We study image classification on CIFAR-10 [12], which contains ten object categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Given a training set $\mathcal{D}_r = \{(x_i, y_i)\}_{i=1}^{N}$ of real images and labels, the goal is to learn a classifier f_θ that predicts the class label y for an input image x. We use the standard 50,000/10,000 train/test split and report top-1 accuracy. All images are treated as RGB inputs at the native CIFAR-10 resolution unless a network requires resizing. The study has two goals. First, we compare a sequence of increasingly expressive recognition pipelines.

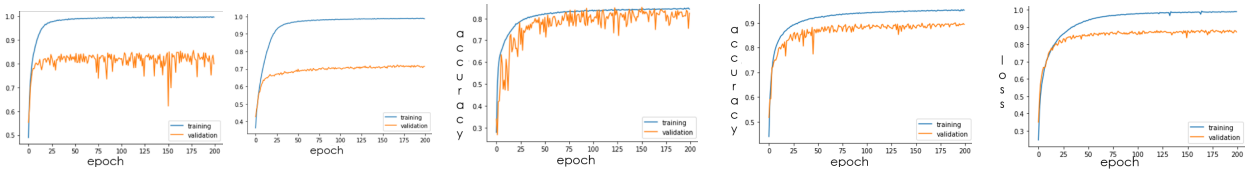

Figure 1. **Training curves of representative deep classification models on CIFAR-10.** We compare the training and validation behavior of VGG19-BN, ResNet, Inception-v4, and DenseNet with transfer learning. The curves show that learned deep representations substantially outperform handcrafted feature baselines, while optimized training schedules and transfer learning lead to improved convergence and better test accuracy.

Second, we use their errors to specify how future generated images should be evaluated.

3.2. Classical Feature-Based Baselines

We first evaluate handcrafted feature representations. For each image, we extract HOG descriptors [4], which summarize local gradient orientation distributions. Following the project setting, each 8×8 pixel cell contributes an orientation histogram, and each 2×2 group of cells is normalized as a block before concatenation into a descriptor. The extracted feature vector is then used as input to classical classifiers:

$$\hat{y} = g(\phi_{\mathrm{HOG}}(x)), \tag{1}$$

where ϕ_{HOG} denotes the HOG feature extractor and g is a classifier such as K-nearest neighbors, support vector machine, or random forest. For KNN, we vary the number of neighbors and compare uniform and distance weighting. For SVM, we use a linear classifier and tune the regularization parameter C. For random forests, we vary the number of estimators. These baselines provide a reference for understanding the limitations of handcrafted features on small natural images.

3.3. Deep Classification Baselines

We then train deep neural networks directly on CIFAR-10. For a neural classifier f_θ, we optimize the standard cross-entropy loss:

$$\mathcal{L}_{\mathrm{ce}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbf{1}[y_i = c] \log p_\theta(c|x_i), \tag{2}$$

where $C = 10$ for CIFAR-10 and $p_\theta(c|x_i)$ is the predicted probability for class c. We compare representative convolutional and transformer-based architectures, including VGG, ResNet, Inception, DenseNet, and Pyramid Vision Transformer. The first group uses a common training recipe from the original project: Adam optimization, categorical cross entropy, batch size 128, dropout before the final fully connected layer, and L_2 regularization. We then compare this recipe with a stronger SGD-based schedule using step decay, motivated by residual-network training practice [7].

3.4. Deep Feature Extraction with Classical Classifiers

To separate feature learning from classifier design, we also use trained neural networks as feature extractors. Given a trained model, we take the activation before the final classification layer as a deep feature:

$$z_i = \phi_\theta(x_i). \tag{3}$$

We then train classical classifiers on z_i, including SVM, random forest, and KNN. This setting evaluates whether the final softmax classifier is optimal or whether classical classifiers can better exploit the learned representation. It also separates two sources of improvement: representation learning in the backbone and decision-boundary learning in the classifier.

3.5. Transfer Learning

We further consider transfer learning using an ImageNet-pretrained DenseNet model. The classifier head is first trained while the backbone is fixed, and then the full network is fine-tuned. This strategy allows the model to benefit from large-scale visual pretraining while adapting to CIFAR-10. Transfer learning serves as a strong baseline because it already incorporates external visual knowledge. In the project report, the top layers are trained first and the full model is then unfrozen for fine-tuning. This two-stage procedure obtains the strongest result in our study.

3.6. Generative Augmentation Protocol

The workshop theme motivates a natural extension: using generated images to improve recognition. Because the available project report does not contain completed generative experiments, we define the protocol explicitly rather than reporting unverified synthetic-data results. Let $\mathcal{D}_s = \{(\tilde{x}_j, \tilde{y}_j)\}_{j=1}^{M}$ be a synthetic training set generated by a class-conditioned model. For each class c, a generative model produces synthetic samples:

$$\tilde{x} \sim p_\psi(x|y = c), \tag{4}$$

where p_ψ denotes a generative image model such as a GAN or diffusion model. The augmented training set is

$$\mathcal{D}_{aug} = \mathcal{D}_r \cup \lambda \mathcal{D}_s, \tag{5}$$

where λ controls the relative contribution of synthetic images during training. We then train classifiers under four settings:

- **Real only**: training only on CIFAR-10 real images.
- **Synthetic only**: training only on generated images.
- **Real + synthetic**: training on the union of real and synthetic images.
- **Real + filtered synthetic**: training on real images and synthetic images filtered by confidence or similarity criteria.

The purpose of this protocol is to determine whether synthetic images improve downstream classification, especially in ambiguous classes. Rather than only reporting average accuracy, we evaluate class-wise confusion and high-confidence failure cases. For a test image x_i, we define a failure margin

$$m_i = \max_{c \neq y_i} p_\theta(c|x_i) - p_\theta(y_i|x_i). \tag{6}$$

Large positive margins identify confident mistakes. Future generative augmentation is considered useful only if it improves accuracy while reducing these margins for ambiguous class pairs.

4. Experiments

4.1. Dataset and Evaluation Metrics

We conduct experiments on CIFAR-10 [12]. The dataset contains 50,000 training images and 10,000 test images from ten object classes. Each image is an RGB image of size 32×32. We report top-1 classification accuracy as the primary metric. To analyze class-level behavior, we also report confusion matrices and inspect high-confidence failure cases. Unless otherwise stated, accuracies are reported in percent. The numbers in this section are transcribed from the project report and reorganized into a common evaluation format.

4.2. Classical Baselines

We first evaluate HOG features combined with KNN, SVM, and random forest classifiers. These methods provide a non-deep-learning baseline for CIFAR-10 classification. The best KNN setting uses distance weighting with $k = 6$ and reaches 47.99% test accuracy. For the linear SVM, increasing C improves performance up to the tested range, with the best reported test accuracy of 50.64%. The random forest baseline reaches 55.13% but obtains 100% training accuracy, indicating severe overfitting. Overall, HOG-based classifiers achieve substantially lower accuracy than deep neural networks. This result suggests that handcrafted gradient statistics are insufficient for capturing semantic variation in CIFAR-10, especially for animal categories and visually ambiguous object classes.

Table 1. Classical baselines using HOG features on CIFAR-10.

Method	Feature	Train Acc.	Test Acc.
KNN	HOG	100.00	47.99
SVM	HOG	50.40	50.64
Random Forest	HOG	100.00	55.13

Table 2. Deep neural network baselines on CIFAR-10.

Model	Training Acc.	Test Acc.
VGG19-BN	85.78	82.64
ResNet	95.10	89.36
Inception-v4	98.87	87.13
PVT-Tiny	98.58	71.43
DenseNet	99.36	78.40
DenseNet Transfer	99.82	93.67

Table 3. Effect of hyperparameter optimization.

Model	Original Test Acc.	Optimized Test Acc.
VGG19-BN	82.64	92.29
ResNet	89.36	93.60
Inception-v4	87.13	89.70

4.3. Deep Network Baselines

We next compare representative deep neural architectures. Compared with HOG-based methods, deep networks achieve significantly stronger performance, indicating the importance of learned hierarchical representations. Among the evaluated models, ResNet and DenseNet provide strong results, while transfer learning with DenseNet achieves the best performance in our preliminary study.

4.4. Effect of Hyperparameter Optimization

We observe that hyperparameter choices have a large effect on deep classification performance. In particular, replacing a fixed learning rate with a step-decay schedule and using stochastic gradient descent improves the performance of several architectures. This observation is consistent with the training strategy used in residual networks [7]. The results highlight that comparisons between architectures should be interpreted carefully unless the optimization settings are sufficiently tuned. The improvement is most visible for VGG19-BN and ResNet. VGG19-BN rises from 82.64% to 92.29% test accuracy in the original report's optimized run, while ResNet reaches 93.60%. Inception-v4 also improves, although less strongly than ResNet under the reported schedule.

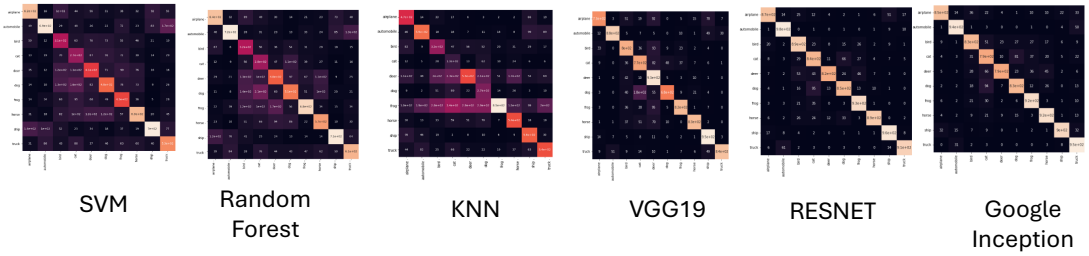

Figure 2. **Confusion matrix of the strongest CIFAR-10 classifier.**

Table 4. Deep feature extraction followed by classical classifiers.

Feature Backbone	SVM	Random Forest	KNN
VGG19-BN	85.63	85.79	84.92
ResNet	90.59	90.91	90.20
Inception-v4	88.17	88.12	88.10

Table 5. Final ranking of the strongest reported settings.

Setting	Test Acc.
DenseNet transfer learning	93.67
Optimized ResNet	93.60
Optimized VGG19-BN	92.29
ResNet features + random forest	90.91
Optimized Inception-v4	89.70
HOG + random forest	55.13

4.5. Deep Features with Classical Classifiers

We also evaluate whether features extracted from trained neural networks can be better classified by classical classifiers. For each trained model, we extract features before the final softmax layer and train SVM, random forest, and KNN classifiers. The results show that deep features combined with classical classifiers can slightly improve over the original softmax classifier in some cases. This suggests that the learned representation contains discriminative information that may not be fully exploited by the final classifier layer. The strongest feature-classifier combination in this setting is ResNet features with a random forest, reaching 90.91% test accuracy. Although the gains over softmax are modest, the result is useful diagnostically because it shows that representation quality and classifier choice can be studied separately.

Fig. 1 shows the training curves of the evaluated deep models. Compared with the original optimization setting, improved learning-rate scheduling leads to more stable convergence and higher test accuracy. Among the evaluated models, DenseNet with transfer learning achieves the strongest final performance. To further understand the failure modes, we visualize the confusion matrix of the strongest model in Fig. 4.3. Although the model achieves high overall accuracy, several class pairs remain difficult to separate. In particular, cat/dog, deer/horse, automobile/truck, and airplane/ship show recurring confusion, indicating that background correlation and semantic similarity still affect the classifier.

4.6. Transfer Learning and Model Ranking

The best result in the report is obtained by DenseNet transfer learning. The model is initialized from ImageNet, trained first through newly added top layers, and then fine-tuned end to end. After the first stage, the model reaches 79.69% test accuracy; after full fine-tuning, it reaches 93.67%. This result is close to the optimized ResNet baseline but slightly higher, making DenseNet transfer learning the strongest reported model. The comparison is important for generative augmentation: synthetic images should be compared against this strong transfer baseline, not only against weaker from-scratch models.

4.7. What These Results Mean for Synthetic Data

The baseline results constrain what a useful generative augmentation study must show. First, generated data should be evaluated against a strong transfer-learning baseline because transfer already imports external visual knowledge. Second, improvements should be class-wise, not only average. For example, if synthetic samples increase the number of airplane images but preserve the same blue-sky background bias, they may not reduce airplane/bird/ship confusion. Third, synthetic data should be filtered when the conditioning label and image semantics disagree. This is especially important for classes with fuzzy boundaries, such as automobile/truck and cat/dog.

5. Error Analysis and Discussion

5.1. Class-Wise Confusion

Overall accuracy provides a useful summary but does not fully explain how models fail. Therefore, we analyze con-

fusion matrices and high-confidence failure cases. The strongest baseline still makes systematic mistakes between visually similar categories. For example, cats and dogs are frequently confused because of similar texture, pose, and facial structure. Deer and horses are also difficult to distinguish when horns are not visible or when the object occupies only a small region of the image. Automobiles and trucks form another ambiguous pair, especially for vans or vehicles captured from partial viewpoints. Airplanes, birds, and ships can be confused when the background is dominated by sky or water.

These errors suggest that CIFAR-10 classification is not only a problem of model capacity. Some errors are caused by low resolution, ambiguous labels, background correlation, and semantic overlap between classes. This observation is important for evaluating generative augmentation. If synthetic images merely increase the number of background-correlated samples, they may not reduce these errors. A useful synthetic augmentation method should instead improve class-specific visual diversity while preserving label consistency.

5.2. High-Confidence Failure Cases

We further inspect high-confidence wrong predictions. A high-confidence failure occurs when the model assigns a high probability to an incorrect class while assigning a low probability to the true class. These cases are particularly informative because they reveal systematic biases rather than random uncertainty. For example, an airplane with a blue background may be predicted as a bird or ship, while a small truck may be predicted as an automobile. Some examples are also visually ambiguous to humans, which indicates that the ground-truth label may not always capture the perceptual ambiguity of the image. The original report also compares the strongest DenseNet transfer model with the second-best optimized ResNet model. This comparison is useful because common errors across two different architectures are less likely to be incidental. For airplanes, many common errors involve small objects against sky or water backgrounds. For automobiles and trucks, the common failures often involve vans or partial vehicle views. For animal classes, the recurring errors are concentrated around cat/dog and deer/horse boundaries, where texture, pose, and missing discriminative parts make the class label uncertain.

This analysis motivates the use of class-wise and failure-aware metrics for synthetic data evaluation. If generative augmentation improves only easy classes while worsening ambiguous classes, its overall benefit may be limited. Conversely, if synthetic images reduce high-confidence errors in ambiguous categories, they may provide meaningful complementary information.

5.3. Architecture-Level Observations

The comparison across architectures gives several practical lessons. First, handcrafted HOG features are too local and edge-oriented to describe the semantic structure of CIFAR-10 objects. They work better for classes with distinctive shape and background patterns, such as airplane, ship, and truck, but struggle on animals. Second, convolutional architectures remain strong on this benchmark. ResNet benefits from residual optimization, while DenseNet benefits from feature reuse and transfer learning. Third, the PVT-Tiny result is weaker than expected from large-scale recognition literature. This does not imply that transformers are unsuitable for CIFAR-10; rather, it suggests that small low-resolution datasets and limited training recipes can make transformer optimization fragile. Fourth, replacing the final softmax classifier with a classical classifier over deep features produces small gains for several backbones. This indicates that the penultimate representation is discriminative, but the final decision layer may not always be the best possible classifier under the chosen training setting.

5.4. Implications for Generative Augmentation

The proposed generative augmentation protocol should be evaluated with three criteria.

Fidelity. Synthetic images should be visually plausible and consistent with the target class. Low-quality images may introduce noise and reduce classifier performance.

Diversity. Synthetic images should cover variations not sufficiently represented in the real training set. If the generated samples are too similar to existing images, they may provide little benefit.

Label consistency. The generated image should match the intended class label. For ambiguous categories such as cat/dog and automobile/truck, label inconsistency can be especially harmful.

These criteria suggest that synthetic data should not be used blindly. Instead, it should be filtered or weighted according to confidence, similarity, or semantic consistency. For example, a pretrained classifier or vision-language model can be used to remove generated samples whose predicted class disagrees with the conditioning label. Feature-space similarity can also be used to reject samples that are too far from the real data distribution. For CIFAR-10, a practical protocol would generate a fixed number of images per class, train the same classifier under real-only and real-plus-synthetic settings, and then report both overall accuracy and per-class confusion changes. The most important comparisons should focus on the ambiguous pairs identified above. If synthetic images improve easy classes but do not improve these pairs, the augmentation is unlikely to address the main recognition failure modes. If the generated images reduce common DenseNet/ResNet errors, the evidence for useful generative augmentation is much stronger.

5.5. Submission Relevance

This study is relevant to a workshop on generative models for computer vision because it frames synthetic data as a recognition problem rather than only a synthesis problem. The central question is not whether a generative model can produce plausible CIFAR-like images, but whether those images contain the variations needed to improve recognition. The baseline and error analysis in this paper provide the measurement scaffold for that question. In a non-archival workshop setting, this makes the work suitable as an ongoing study: the current paper establishes the recognition baselines, while the proposed protocol defines the next set of generative experiments.

5.6. Limitations

This study has several limitations. First, CIFAR-10 is a small and low-resolution dataset, so conclusions may not directly transfer to large-scale or high-resolution recognition tasks. Second, the current baseline study focuses on classification accuracy and error analysis; the full generative augmentation experiments require additional synthetic data generation and filtering. Third, different generative models may produce synthetic images with different levels of fidelity and diversity. Therefore, future work should compare GAN-based, diffusion-based, and retrieval-augmented synthetic data sources under the same evaluation protocol.

6. Conclusion

We presented a preliminary CIFAR-10 study for analyzing small-scale visual classification with classical features, deep networks, transfer learning, and generative augmentation. Our baseline experiments show that deep learned representations substantially outperform handcrafted HOG features, and that DenseNet-based transfer learning provides the strongest performance among the evaluated models. More importantly, our error analysis reveals persistent class ambiguities that are not fully explained by average accuracy alone. These findings motivate a generative augmentation protocol that evaluates synthetic images not only by their effect on overall classification accuracy, but also by their impact on class-wise confusion and high-confidence failure cases. We hope this study provides a simple and controlled framework for understanding when synthetic images help visual recognition and when they may introduce additional semantic or background bias.

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. In *Transactions on Machine Learning Research*, 2023. 1, 2

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1): 5–32, 2001. 1, 2

[3] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. 1, 2

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005. 1, 2, 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1, 2

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 1, 2

[7] 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. 1, 2, 3, 4

[8] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip H. S. Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 1, 2

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 1, 2

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 1, 2

[11] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[12] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 1, 2, 4

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2

[14] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8011–8021, 2023. 1, 2

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 1, 2

[17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 1, 2

[18] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 1, 2

[19] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, 2014. 2

[20] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 2