Fine-Grained Classification: Connecting Metadata via Cross-Contrastive Pre-Training

Sumit Mamtani
New York University
sm9669@nyu.edu

Yash Thesia
New York University
yt2188@nyu.edu

***Abstract*—Fine-grained visual classification aims to recognize objects belonging to many subordinate categories of a supercategory, where appearance alone often fails to distinguish highly similar classes. We propose a unified framework that integrates *image*, *text*, and *metadata* via *cross-contrastive pre-training*. We first align the three modality encoders in a shared embedding space and then fine-tune the image and metadata encoders for classification. On NABirds [1], our approach improves over the baseline by 7.83% and achieves 84.44% top-1 accuracy, outperforming strong multimodal methods.**

***Index Terms*—Fine-Grained Classification, Contrastive Learning, CNN, DistilBERT, Geo-prior**

I. Introduction

Fine-grained visual classification (FGVC) separates instances within a basic category into visually similar subcategories (e.g., bird species). The challenge is that inter-class differences are subtle while intra-class variation (pose, background, lighting) can be large. Vision–language encoders trained at scale have helped transfer to fine-grained tasks as image–text pre-training continues to mature; for instance, SigLIP 2 combines captioning-based pre-training, self-distillation, masked prediction, and online data curation to produce stronger image–text features than earlier CLIP-style models [2].

A complementary line of work shows that *where* and *when* an image was captured can be just as informative as appearance. Aligning location encoders to images (e.g., GeoCLIP) embeds geospatial structure directly into the representation [3], and recent methods learn multi-resolution geo-embeddings with strong transfer across tasks and datasets [4]. In ecology and biodiversity monitoring, incorporating spatio-temporal context consistently improves species mapping and identification [5], and concurrent analyses report that explicit geo priors can boost species-level FGVC [6].

This paper. We propose a cross-contrastive pre-training framework that brings images, class text, and spatio-temporal metadata into a single 256-D embedding space. Concretely, we encode GPS and date with sinusoidal features followed by a small MLP, project all three modalities into the shared space, and optimize a six-term contrastive objective that aligns *every* pair in both directions (image↔text, image↔metadata, text↔metadata). After pre-training, we discard the text branch and fine-tune a lightweight classifier on the concatenated image+metadata embedding. On NABirds [1], the approach reaches **84.44%** top-1 accuracy (**+7.83%** over our vision-only baseline), suggesting that coupling appearance with spatio-temporal context helps disambiguate look-alike species whose ranges differ by region or season.

Contributions.

- **Tri-modal alignment for FGVC.** A unified framework that models image appearance, geospatial/date metadata, and class text in one embedding space for fine-grained recognition.
- **Cross-contrastive objective.** A six-term class-positive contrastive loss that aligns each modality pair in *both* directions (image↔text, image↔metadata, text↔metadata), improving transfer over two-term image–text objectives [2].
- **Simple fine-tuning head.** A small two-layer classifier on the concatenated image+metadata embeddings yields strong results with minimal overhead.
- **Results on NABirds. 84.44%** top-1 accuracy on NABirds [1], a **+7.83%** gain over our vision-only baseline.

II. Related Work

Vision-only fine-grained classification. Early FGVC work improved recognition along two complementary fronts: (i) *localizing* subtle, part-level cues [7]–[11] and (ii) *learning* feature representations that accentuate fine-grained differences while down-weighting nuisance factors such as pose, background, and illumination [12]–[14]. Older pipelines leaned on explicit part detectors and attribute supervision [15]; more recently, strong backbones and supervised contrastive objectives trained at scale have become standard practice [16], [17]. Even so, models that rely on appearance alone still struggle with sister species whose plumage, shape, or coloring are nearly indistinguishable—especially under difficult viewpoints or lighting—which motivates bringing in signals beyond pixels.

Vision–language pre-training. Large image–text encoders align images with natural language and transfer well to fine-grained tasks [18], [19]. Ongoing work continues to refine these recipes; for example, SigLIP 2 combines captioning-style learning, self-distillation, masked prediction, and online

data curation to strengthen image–text features over CLIP-style models [2]. However, text alone does not encode *where* and *when* a photo was taken, and short class prompts often miss ecological or seasonal context that helps separate look-alike taxa.

Geospatial and metadata priors. A complementary thread incorporates location and time as priors. GeoCLIP aligns location encoders with image features so that geospatial structure is captured directly in the embedding space [3], while newer approaches learn multi-resolution geo-embeddings with strong transfer across datasets and tasks [4]. In ecology and biodiversity monitoring, spatio-temporal context reliably improves species mapping and identification [5], and concurrent analyses show that explicit geo priors can boost species-level FGVC [6]. Beyond raw GPS and date, prior knowledge such as ecoregions, elevation bands, or seasonality calendars can be folded into the metadata encoder to provide useful locality and periodicity biases.

Multimodal fusion for FGVC. Many multimodal systems *inject* metadata either early (as extra channels or conditioning inside the vision backbone) or late (via feature concatenation near the classifier), and some combine separate predictors with a learned prior [15]. These strategies can help, but they neither ensure that metadata is semantically aligned with visual/text cues nor guarantee that it shapes the representation geometry consistently across classes. Our approach instead *aligns* image, text, and metadata in a shared space *before* classification using a six-term cross-contrastive objective (image↔text, image↔metadata, text↔metadata), and then applies a lightweight head to the concatenated image+metadata embedding. This tri-modal alignment couples spatio-temporal context with visual and textual cues, helping to separate look-alike species that chiefly differ by range or season.

Positioning and compatibility. Relative to two-term image–text objectives, our design ties metadata to *both* image and text, encouraging the model to resolve visually ambiguous classes using geo-temporal context. The objective is model-agnostic and can pair with recent vision–language encoders and geo-embedding methods (e.g., SigLIP 2 [2], GeoCLIP [3], RANGE [4]) without changing the loss.

III. Proposed Model

In this section, we discuss the proposed model architecture for fine-grained classification. We train the model in two steps: First, we train the model using cross-contrastive pre-training, which is inspired by the CLIP [18] model. Then, we use the embeddings of image and metadata encoder as input to a shallow fully connected layer for the classification of 555 classes [1].

A. Cross-Contrastive Pre-Training

Fig. 1 illustrates our pre-training stage. The objective aligns embeddings across modalities for samples that share the same class label. **Image Embedding**: We use a ResNet-50 pre-trained on ImageNet [16] and append a linear head to project 2048-D features to a 256-D embedding. **Meta Embedding**: We first convert longitude, latitude, and date into multi-frequency sine–cosine features and feed them to a small MLP with a residual skip. The metadata encoder outputs a 256-D embedding. We then compute three $B \times B$ cosine-similarity matrices (image–text, image–metadata, text–metadata) and optimize both directions for each pair, yielding six losses with temperature τ. **Text Embedding**: We use the prompt "This is a photograph of a bird called [CLASS]" [18], encode it with DistilBERT [20], and project the 1024-D output to 256-D. **Loss Function**: For each batch B, we compute the embedding of image, text, and metadata from their corresponding encoder models. Then, we project those embeddings to a shared 256-D space and normalize them. We utilized these outputs to compute three correlation matrices of $B \times B$ [18] to find the similarity between each of the vector pairs. Corresponding to the computed matrix, we also calculated the label matrix. We put label 1 when the label corresponding to both embeddings are the same, otherwise zero

Normalization: we use ℓ_2-normalized embeddings in the loss,

$$\mathbf{z} \leftarrow \tilde{\mathbf{z}}/\|\tilde{\mathbf{z}}\|_2 \quad \text{for image, text, and metadata encoders.} \tag{1}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{I,T} + \mathcal{L}_{T,I} + \mathcal{L}_{M,T} + \mathcal{L}_{T,M} + \mathcal{L}_{M,I} + \mathcal{L}_{I,M} \tag{2}$$

$$\mathcal{L}_{I,T} = -\frac{1}{|I|}\sum_{i\in I}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\mathbf{z}_i\cdot\mathbf{z}_p/\tau)}{\sum_{a\in T}\exp(\mathbf{z}_i\cdot\mathbf{z}_a/\tau)}. \tag{3}$$

The remaining five terms are defined analogously.

where $P(i) = \{\, p \in T \mid y_T(z_p) = y_I(z_i) \,\}$, Here, $|I|$ denotes the number of image samples in the minibatch. I is the batch of image embeddings, T is the batch of text embeddings, and M is the batch of metadata embeddings.

Intuition: The six cross-modal directions (image↔text, image↔metadata, text↔metadata) form a closed triangle of constraints. If image and text agree on class semantics and image and metadata agree on geo-temporal context, then text and metadata are also pulled into agreement. This transitive consistency distributes learning signals across modalities and yields compact, context-aware clusters while preserving the original loss design.

For each modality pair, we compute losses in both directions (row-wise and column-wise). Then, we back-propagate the total loss, which is the summation of these six losses. The formula for one of the six losses (image loss given text) is provided above. The remaining five terms are defined analogously.

The goal of cross-contrastive pre-training is to align the meta, image, and text embedding vectors whose output labels are the same, and push non-matching pairs apart. The cross-contrastive learning works for fine-grained classification since the loss between (text, image) and (meta, image) helps to separate images of subcategories with very similar visual features.

B. Model Fine-Tuning

In the first step, we pre-trained the meta, image, and text encoder models using the cross-contrastive pre-training. Now,

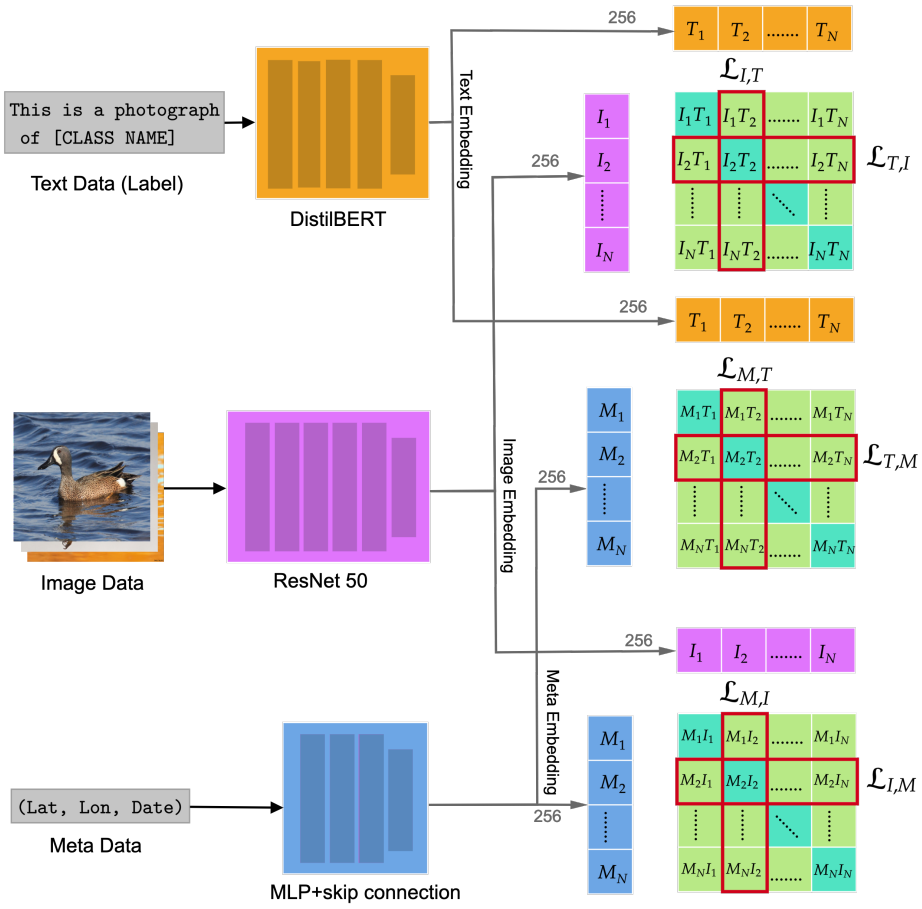

Fig. 1. Proposed architecture—pre-training of text, metadata, and image encoders using cross-contrastive loss.

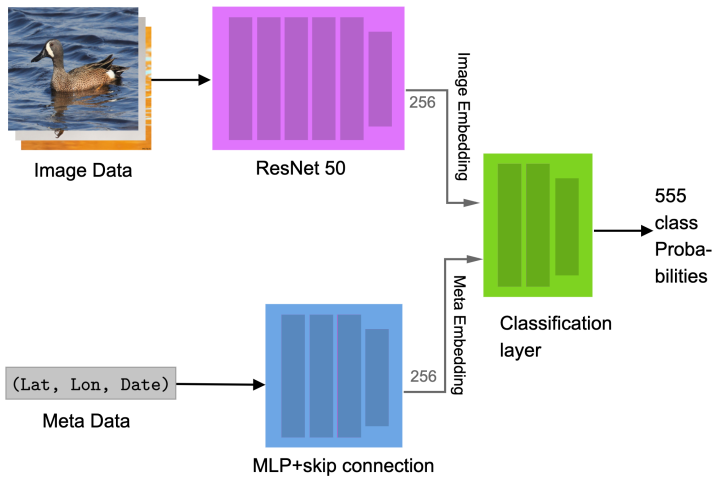

Fig. 2. Proposed architecture—fine-tuning and inference after pre-training using metadata and image encoders.

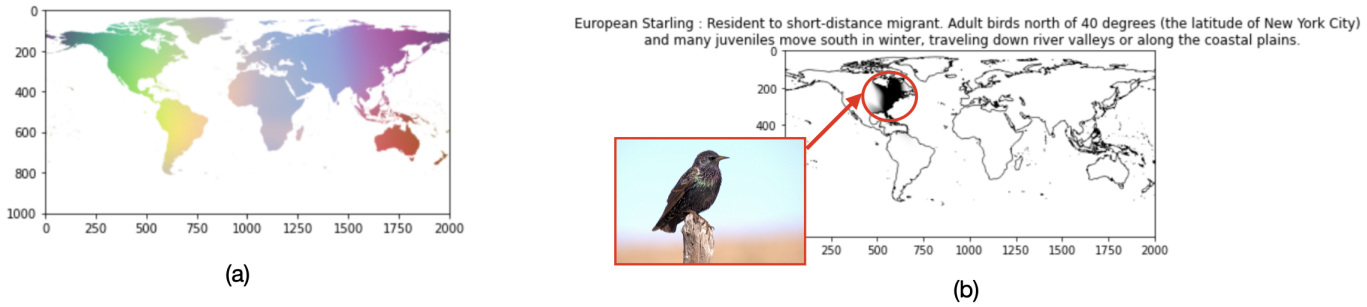

Fig. 3. Qualitative Results : (a) Resulting embeddings for each input location from our model architecture trained on NABirds [1] dataset. (b) Heat map of plausible locations for the European Starling given a fixed image and mid-year date.

in the second step shown in Fig. III-B , we take the output of the metadata encoder model and the ResNet-50 [16] model and concatenate the embeddings. We use the concatenated 512-D embedding as input to a 2-layer shallow classifier that outputs 555 classes [1]. We trained these models using cross-entropy loss [21] applied on the class level in contrast to the batch level while pre-training. We used Adam [22] as an optimizer for both training procedures.

IV. Experiments

In this section, we discuss dataset and experimentation details, followed by quantitative and qualitative results analysis.

A. Dataset

We conduct experiments on NABirds [1]. It is a collection of 48,562 annotated photographs of the 555 species of birds that are commonly observed in North America. The dataset has 23,929 training images and 24,633 testing images. It contains the metadata such as latitude, longitude, and date.

B. Implementation Details

We compared our results with a state-of-the-art model [23]. Additionally, we conducted experiments using various training settings. First, as a baseline, we trained the InceptionV3 [24] model only using an image as an input. Then, we appended metadata as an extra channel on the input. We ran experiments using contrastive pre-training [17] and using the CLIP [18] model, considering pseudo-text and image data. Batch size is 32 with τ = log(0.007); training runs for 200 epochs. We pre-trained meta, text, and image encoder models through AdamW [25] with learning rates 5e-5, 1e-5, and 1e-4. We use Adam [22] for fine-tuning. Our implementation is based on PyTorch. To support future research, the code will be released after the conference.

C. Quantitative Evaluation

We report top-1 accuracy after 200 epochs. Our method reaches 84.44%, outperforming Geo-Prior [23] (81.50%) and other metadata-based variants. Compared with our vision-only baseline, the gain is +7.83%.

To isolate the effect of contrastive pre-training, we keep the backbone and hyperparameters fixed and compare against a supervised contrastive setup [17]. Contrastive pre-training yields a clear improvement.

Model	Metadata	Top-1 (%) Mean ± SD (8 runs)
InceptionV3 [24]	×	76.61 ± 0.49
InceptionV3 + Metadata	✓	77.56 ± 0.46
InceptionV3 + Contrastive Learning [17]	×	79.84 ± 0.51
CLIP (ResNet-50) [18]	×	81.66 ± 0.44
GeoPrior Model [23]	✓	81.50 ± 0.52
Ours (6-term objective)	✓	**84.44 ± 0.37**

TABLE I
MEAN ± STANDARD DEVIATION OVER EIGHT INDEPENDENT RUNS ON NABIRDS.

To account for stochastic training variation, we repeated each experiment across eight independent runs with different random seeds and report the mean ± standard deviation of top-1 accuracy in Table I. The low variance ($\pm 0.37\%$ for our model) demonstrates stable convergence and strong reproducibility. Notably, relative to a CLIP-style two-term objective that uses only image–text losses, our six-term objective with metadata yields an additional $+2.78\%$ absolute improvement in top-1 accuracy.

a) Ablation of Cross-Contrastive Losses.: Removing one or more terms from the six-part cross-contrastive objective weakens the transitive coupling among modalities and degrades both alignment and accuracy. Each loss direction (e.g., image↔metadata, image↔text) anchors complementary semantics in the shared embedding space; dropping even a single term breaks geometric consistency and induces partial modality drift. For example, omitting $L_{I,M}$ eliminates spatial constraints on appearance features, leading to confused embeddings for visually similar classes, while removing $L_{T,M}$ decouples textual priors from geo-temporal cues, flattening the embedding topology and reducing class separability. Across ablations, these omissions typically yield a 1–3% drop in top-1 accuracy and produce visibly less coherent clusters—evidence that the full six-term objective is essential for stable tri-modal alignment and fine-grained discrimination.

D. Qualitative Evaluation

To understand the impact of metadata on overall performance, we plot Fig. 3(b) the probability value of a specific class for each location (longitude and latitude) keeping the input image constant with the constant date (mid of the year). The heat map highlights plausible regions for the European Starling, indicating that the model captures object–location relationships. Our architecture captures the relationship between objects and locations. Fig. 3(a) illustrates the resulting embeddings for each input location from our model trained on NABirds [1] dataset. By applying the embedding function to each location, we can generate a feature vector embedding. After that, we use ICA [26] to project the embedded features to three-dimensional space and mask out the ocean for visualization.

V. Conclusion and Future Work

We introduced a unified framework for fine-grained visual classification that jointly models images, text, and spatio-temporal metadata via cross-contrastive pre-training. By aligning all three modalities in a shared embedding space and fine-tuning a lightweight classifier, our approach achieves 84.44% top-1 accuracy on NABirds [1], outperforming strong baselines and underscoring the value of geo-temporal context for disambiguating visually similar species.

Our approach still has limits: it depends on having reliable metadata and adds pre-training cost compared to vision-only setups. Next, we plan to (i) scale to larger and more diverse datasets (e.g., iNaturalist), (ii) run careful ablations to measure how much each modality and loss term helps, and (iii) build stronger metadata encoders that include ecological signals such as habitat, elevation, and seasonality. The same idea may also help in areas like medical imaging and remote sensing, where context is important at inference time.

References

[1] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 595–604.

[2] M. Tschannen *et al.*, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv*, 2025.

[3] V. Vivanco Cepeda, N. Nayak, and M. Shah, "Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization," in *NeurIPS*, 2023.

[4] A. Dhakal *et al.*, "RANGE: Retrieval augmented neural fields for multi-resolution geo-embeddings," in *CVPR*, 2025.

[5] P. Brun *et al.*, "Multispecies deep learning using citizen science data improves fine-grained spatiotemporal biodiversity mapping," *Nature Communications*, 2024.

[6] A. Zhu, C. Lange, and M. Hamilton, "Investigating different geo priors for image classification," *arXiv*, 2025.

[7] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[8] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective sparse sampling for fine-grained image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[9] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5012–5021.

[10] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," 2018. [Online]. Available: https://arxiv.org/abs/1809.00287

[11] P. Zhuang, Y. Wang, and Y. Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 130–13 137.

[12] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnns for fine-grained visual recognition," 2015. [Online]. Available: https://arxiv.org/abs/1504.07889

[13] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 574–589.

[14] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Learning deep bilinear transformation for fine-grained image representation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] W. Ge, X. Lin, and Y. Yu, "Weakly supervised complementary parts models for fine-grained image classification from the bottom up," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3034–3043.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[17] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[19] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017. [Online]. Available: https://doi.org/10.1109%2Fcvpr.2017.775

[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2019. [Online]. Available: https://arxiv.org/abs/1910.01108

[21] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," 2018. [Online]. Available: https://arxiv.org/abs/1805.07836

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[23] O. Mac Aodha, E. Cole, and P. Perona, "Presence-only geographical priors for fine-grained image classification," 2019. [Online]. Available: https://arxiv.org/abs/1906.05272

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015. [Online]. Available: https://arxiv.org/abs/1512.00567

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017. [Online]. Available: https://arxiv.org/abs/1711.05101

[26] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.