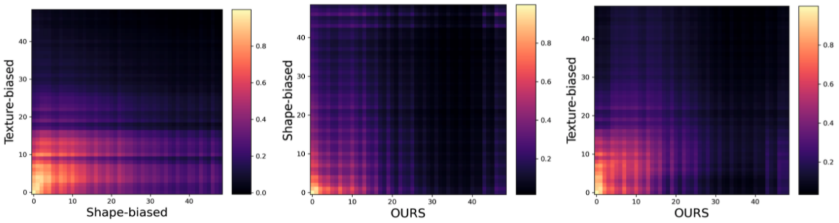
A Description on Baseline Method

To briefly describe this baseline, given a set of samples, it first creates shape and texture pairs. The shape pair implies a pair of samples that share similar shape information and distinct texture characteristics, whereas the texture pair includes two samples under similar texture information and different shape characteristics. To create texture pairs, the baseline model utilizes heuristically-chosen patterns (i.e., cross-shape patterns) as target textures and trains an additional Style-transfer model based on these patterns. Given texture and shape pairs, the baseline then extracts representation vectors (denoted as Z) from the trained classifier. It then calculates the mutual information score of a given pair following the equation 1, and establishes a set of mutual scores as follows: shape mutual score, texture mutual score, and residuals. Note that residuals are fixed as 1 in the original publication of Islam et al. [4]. Lastly, they additionally define the final bias score by multiplying the dimensions of the CNN's last layer with the softmax-activated mutual scores. Therefore, the baseline yields the number of texture-associated and shape-associated neurons, and regard a particular CNN has high texture bias if the number of texture-associated neurons exceeds the other one. Please refer to the original publications [4] for more detailed elaborations on the baseline approach.

$$MI(Z_i^a, Z_i^b) > -\frac{1}{2}log(1-\tau_i^2); \tau_i = \frac{Cov(Z_i^a, Z_i^b)}{\sqrt{Var(Z_i^a)Var(Z_i^b)}}. \qquad (1)$$



(a) Shape v. Texture (b) Shape v. Granular (c) Texture v. Granular

Fig. 1. Representation similarities among layers at different models. The model trained under the granular labeling scheme acquires less similarity with the one trained under shape and texture schemes. We presume the granular labeling scheme contributes to a more qualified high-level representation of the model, which is known to describe contextual understanding on a given sample. Note that S-biased means shape-biased and T-biased means texture-biased.

B Does Granular Labeling Scheme Yield Different Representation?

Furthermore, we check the representation similarity between different models under the three pairs: (granular, shape-biased), (granular, texture-biased), and (shape-biased, texture-biased). By measuring the similarity between different models, we aim to scrutinize whether the proposed granular labeling scheme learns a knowledge different from the other schemes. Upon the representation similarities between different models shown in Figure 1, we also discovered that the proposed granular scheme acquires a different representation than the others. While the lower-layers (approx. 0 to 18) at the texture-biased model share similar representations with the shape-biased one, our granular model particularly share fewer representations with the others. Especially, the proposed granular labeling scheme contributes to acquiring different representations at high-level layers of the CNN, which are known to implicit contextual, semantic understanding of a given data [6,7]. Considering this results with the analysis proposed in the manuscript, we figured out the effectiveness of the proposed granular labeling scheme comes from the quality of representation. The representation trained under the proposed scheme has a larger knowledge capacity, and it acquires a presumably qualified contextual understanding of a given data at high-level layers of the neural networks.

C Visual Elaborations

For ease of understanding, we provided several visualized examples of the dataset and classification tasks. In Figure 2, we visualized style-transferred samples from the original CIFAR-10 sample. In Figure 3, we illustrated various classification tasks utilized in the study: shape classification, texture classification, and granular classification.

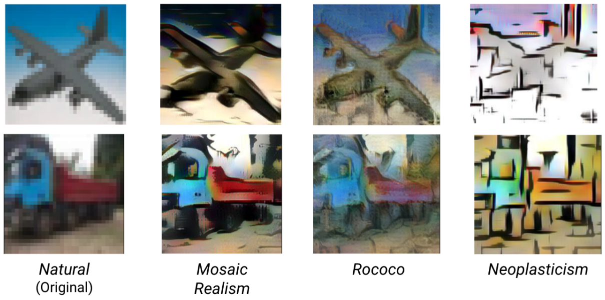
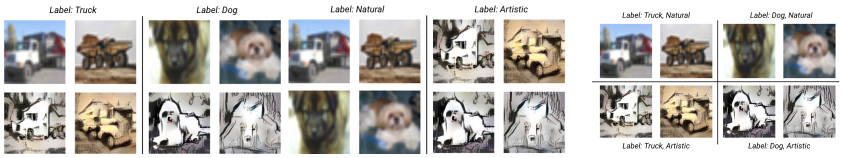


Fig. 2. Example of style-transferred samples from CIFAR-10

D Does Granular Labeling Scheme Contribute to Better Transferability?

D.1 Setup

We further scrutinized whether the proposed granular labeling scheme contributes to better transferability. Given three models' weights trained under



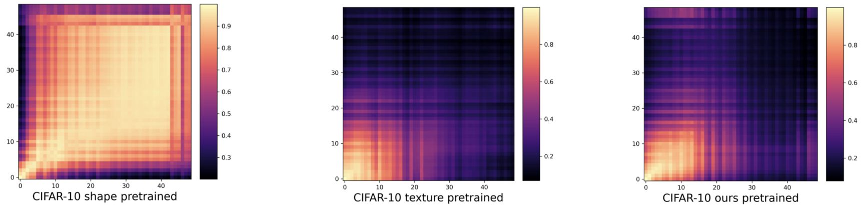
(a) Shape Classification (b) Texture Classification (c) Granular Classification

Fig. 3. Example classification tasks under the concatenated dataset. The shape classification in (a) aims to establish a discriminative boundary based on the object shapes, while the texture classification in (b) solves a binary classification between natural and artistic textures.

different labeling schemes, we utilized them as a pre-trained weight to solve the other downstream tasks. We first initialized the model's parameters with a selected pre-trained weight, trained the model with the training set of the downstream dataset, and measured the classification accuracy and F1-score at the downstream dataset's test set. Then, we set these downstream task accuracy and F1-score as a proxy of the pre-trained model's transferability. We regard the higher downstream task performance implies better transferability as the pre-trained weights would influence the task performance. We employed two datasets of CIFAR-100 [5] and SVHN [2] as a seed of the downstream dataset. For more various problem settings, we additionally style-transferred these original downstream datasets with AdaIN. Given the original and style-transferred downstream datasets, we configured three downstream classification tasks following various labeling schemes: shape classification under the shape-biased scheme, texture classification under the texture-biased scheme, and graunular classification under the granular scheme. Then, we can examine the fine-tuning performances in three downstream classifications when we utilized the pre-trained weight acquired under the various labeling schemes. Upon the aforementioned setup on the downstream classification, we described the transferability of various labeling schemes in Table 1.

D.2 Analogy

Following the experiment results, we discovered the granular labeling scheme was not always effective in every downstream classification task. We also figured out that the shape-biased labeling scheme is very effective in shape classification, and the texture-biased scheme is comparatively influential in texture classification. While the granular labeling scheme accomplished the best target classification performance in every problem setting, it became less competent in transfer learning. To scrutinize an underlying reason behind this phenomenon, we focused on a key difference between the target and downstream classifications, and one primary difference is data distribution. While the target classification deals with the samples in the same distribution as the training set, but the downstream classification does not. We hypothesize that the representation trained under the granular scheme is overly fit to understand the training samples; thus, it lacks transferability into other samples that do not share similar characteristics. We further analyzed representation similarity between the pre-trained and fine-tuned models to examine our hypothesis. The visualized representation similarities are shown in Figure 4.



(a) Shape-biased pre-trained (b) Texture-biased pre-trained (c) Granular pre-trained

Fig. 4. Representation similarities between the pre-trained weights (at x-axis) and fine-tuned weights (at y-axis). While the representations of the pre-trained model under the shape-biased scheme are re-used in the fine-tuned models, the other pre-trained weights are not.

Table 1. Downstream shape, texture and granular classification performance of fine-tuned models based on various pre-trained weights, where these weigths are acquired under various labeling schemes.

Downstream Dataset	Labeling Scheme	Training Set					
		Stylized Set 1		Stylized Set 2		Stylized Set 3	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
		Shape Classification					
CIFAR-100	Shape-biased	**0.4179**	**0.5652**	**0.4230**	**0.5710**	**0.4127**	**0.5690**
	Texture-biased	0.3915	0.5394	0.3903	0.5379	0.3917	0.5386
	Granular	0.3921	0.5393	0.3939	0.5397	0.3916	0.5369
SVHN	Shape-biased	**0.8825**	**0.8823**	**0.8828**	**0.8826**	**0.8834**	**0.8832**
	Texture-biased	0.8797	0.8794	0.8785	0.8784	0.8801	0.8799
	Granular	0.8628	0.8625	0.8683	0.8679	0.8671	0.8667
		Texture Classification					
CIFAR-100	Shape-biased	0.9185	0.9523	0.9163	0.9526	**0.9474**	0.9714
	Texture-biased	**0.9805**	**0.9896**	**0.9731**	**0.9857**	**0.9757**	**0.9871**
	Granular	0.9479	0.9719	0.9505	0.9731	0.9463	0.9696
SVHN	Shape-biased	0.8774	0.9322	0.8657	0.9266	0.8611	0.9231
	Texture-biased	**0.9475**	**0.9727**	**0.9151**	**0.9551**	0.8833	0.9374
	Granular	0.9220	0.9585	0.9101	0.9521	**0.9302**	**0.9632**
		Ganular Classification					
CIFAR-100	Shape-biased	**0.1536**	**0.2504**	**0.1458**	**0.2425**	**0.1540**	**0.2305**
	Texture-biased	0.1509	0.2475	0.1356	0.2236	0.1164	0.2016
	Granular	0.1283	0.2154	0.1263	0.2369	0.1358	0.2223
SVHN	Shape-biased	**0.6558**	**0.7363**	0.6436	0.7218	0.6642	0.7441
	Texture-biased	0.6494	0.7350	0.6295	0.7187	0.6334	0.7268
	Granular	0.6489	0.7207	**0.6438**	**0.7231**	**0.6791**	**0.7564**

Given the representation similarities between the pre-trained and fine-tuned models, we discovered a pre-trained weight under the granular labeling scheme does not bear much similar representation with the fine-tuned models. Following prior studies on effective transfer learning [6], feature re-use is one of the significant factors of good transferability. Accordingly, a well-transferable model would have a particular amount of similar representation with the fine-tuned models if the features in pre-trained models would have been re-used. Unfortunately, the representations at the pre-trained model under the granular scheme could not be excessively re-used during the fine-tuning; therefore, it failed to accomplish good transferability. Conversely, we figured out a model pre-trained under the shape-biased scheme has large similarities with various fine-tuned models; and this re-used features would have contributed to better downstream classification performances. In a nutshell, we presume the proposed granular scheme does not always achieve good transferability in downstream tasks as it overly fits the training set. Still, we acknowledge that our analysis is at a hypothetical level; thus, a more strict analogy on this phenomenon is required. We leave this as an improvement avenue.

E Implementation Details

For implementation details (i.e., dataset, codes), please refer to the attached `https://github.com/socar-esther/Optimal_labeling_scheme`. We utilized conventional cross-entropy loss for the learning objective. Note that we did not utilize any data augmentation techniques as they might influence the representation power of the CNN. For a CNN trained with Style-transferred samples (which is proposed in the previous work [1]), we style-transferred TinyImageNet samples with AdaIn [3], and trained the CNN with the aforementioned configurations.

References

1. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
2. Goodfellow, I.J., Bulatov, Y., Ibarz, J., Arnoud, S., Shet, V.: Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082 (2013)
3. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
4. Islam, M.A., Kowal, M., Esser, P., Jia, S., Ommer, B., Derpanis, K.G., Bruce, N.: Shape or texture: Understanding discriminative features in cnns. arXiv preprint arXiv:2101.11604 (2021)
5. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep. (2009)
6. Raghu, A., Raghu, M., Bengio, S., Vinyals, O.: Rapid learning or feature reuse? towards understanding the effectiveness of maml. arXiv preprint arXiv:1909.09157 (2019)
7. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems **32** (2019)