6. Appendix

6.1. Datasests

ImageNet-100 In this paper, we utilized ImageNet-100 for training the ResNet-50 backbone and for measuring the accuracy of trained models. ImageNet-100 is the subset of ImageNet dataset which contains images of 100 classes from the original dataset. It consists of total 130,000 images for training set (1,300 images per class) and total 5,000 images for validation set (50 images per class).

Counterfactual Imageset. By utilizing CGN [18], we generated total 65,000 counterfactual images whose texture in foreground and background of the object are respectively changed to texture of other ImageNet-100 classes. We generated total 2 sets of counterfactual images in Fig. 3. Fig. 3 (a) shows a set of counterfactual images whose texture in inside and outsied of the object is differently changed to texture of other classes. Fig. 3 (b) shows the other set of counterfactual images whose texture only on backgroud is changed.

OOD Benchmark. The OOD benchmark dataset is proposed by Geirhos *et al*. [6] to measure the model's robustness to texture. This dataset consists of images which 17 kinds of adjustments are applied to its texture. We divided images from the dataset into 2 groups; `OOD-noise` and `OOD-style`. The types of modifications included in `OOD-noise` are *grayscale, contrast, high-pass, low-pass, phase noise, power equalisation, opponent colour, rotation, eidolon I, eidolon II, eidolon III, uniform noise*. `OOD-style` includes the images that the following changes are applied to; *sketch, stylized, edge, silhouette, cue conflict*. We sorted out total of 543 images from `OOD-noise` and total of 200 images from `OOD-style` by following the classes of ImageNet-100.

6.2. Augmentations

A total of 5 types of data augmentations are used in our method. We used `RandomResizedCrop` and `RandomHorizontalFlip` augmentations from PyTorch library's implementation. We also used `ColorJitter`, `RandomGrayscale` and `RandomGaussianBlur` augmentations from Kornia [14] library's implementation. For the details, `ColorJitter` is applied with a probability of 0.8, jitter strength of 0.8 for brightness, contrast, saturation, respectively, and jitter strength of 0.2 for hue. `RandomHorizontalFlip` and `RandomGaussianBlur` are applied with the probability of 0.5 and `RandomGrayscale` is applied with the probability of 0.2.

6.3. Training details

Experiments on the supervised learning method. We pretrained ResNet-50 model on ImageNet-100 for 200 epochs using Adam optimizer with an weight decay of 0.0001, an initial learning rate of 0.001, and batch size of 256. We used cosine annealing to adjust learning rate during pretraining time. We build the model with a common backbone and added 3 multiple heads by following CGN's implementation [18]. We trained the model for 45 epochs respectively on 2 sets of counterfactual images in 3 along with ImageNet-100, using Adam with a weight decay of 0.0001, batch size of 256, and an initial learning rate of 0.001 which is decayed by a factor of 10 at epochs 15, 30. For the model trained on counterfactual image sets, we used 65,000 images from ImageNet-100 and 65,000 images from counterfactual image sets. For counterfactual images which have a total of 3 labels respectively for shape, foreground, and background, each head is given its respective label when training on it and output the logit for each respective label. Except for counterfactual images, the logit from each of the heads is averaged. For the model with shape-focused augmentation, we only trained it on ImageNet-100.

Experiments on contrastive learning methods. We pretrained contrastive learning models with ResNet-50 backbone on ImageNet-100 dataset for 350 epochs using SGD optimizer with an weight decay of 0.0001, an initial learning rate of 0.3, and batch size of 512. We used layer-wise adaptive rate scaling to adjust the learning rate during the pretraining time. For the comparison, we also pretrained contrastive models jointly on ImageNet-100 and counterfactual images on the same hyperparameter setting with the vanilla model. In this case, we pretrained the model jointly on 65,000 images from ImageNet-100 and 65,000 counterfactual images. For the vanilla contrastive learning models and contrastive learning models jointly trained on ImageNet-100 and counterfactual images, we applied the same kinds of data augmentations described in Sec. 6.2. For the contrastive learning models with the shape-focused augmentation scheme, the types of data augmentations are the same with the other cases but we followed our data augmentation scheme described in Figure 2 and its overall application process is in Figure 4. By following linear evaluation protocols in Supervised Contrastive Learning and SimCLR methods, we trained a linear classifier on ImageNet-100 dataset for 100 epochs on top of frozen pretrained models, using SGD optimizer with a batch size of 512 and an initial learning rate of 1.0 which is decayed by a factor of 10 at epochs 60, 80.