A. Appendix

A.1. Methodology

The algorithm for the proposed LSFSL approach is provided in Algorithm 1. The sequential distillation and online self-distillation approach to incorporate shape information is illustrated in Algorithm 2 and Algorithm 3.

Algorithm 2 LSFSL-Distill: Training Algorithm

Input: dataset $\mathcal{D}$, fixed LSFSL-trained RIN model R_t with feature extractor f_t and classifier g_t, randomly initialized student RIN model R_s with feature extractor f_t and classifier g_t, epochs E, softmax operator σ, cross-entropy loss CE, Kullback-Leibler divergence loss KLD, cross-entropy loss factor α, teacher-student decision alignment loss factor β

for epoch $e \in \{1, 2, .., E\}$ **do**
    sample a mini-batch $(x, y) \sim \mathcal{D}$
    $R_t(x) = g_t(f_t(x))$
    $R_s(x) = g_s(f_s(x))$
    $\mathcal{L}_{CER} = CE(R_s(x), y)$
    $\mathcal{L}_{DA} = KLD(R_t(x), R_s(x))$
    $\mathcal{L} = \alpha\mathcal{L}_{CER} + \beta\mathcal{L}_{DA}$
    Update parameters of R_s based on $\mathcal{L}$ using Stochastic Gradient Descent (SGD)
end for
return RIN student model R_s

A.2. Analysis Visualizations

The texture bias, spurious correlation, and statistical regularity analysis are performed by applying different textures by stylization, class-specific tints, and radial low-pass Fourier filters at increasing severities as shown in Figure 5, Figure 6 and Figure 7.

Algorithm 3 LSFSL-Online: Training Algorithm

Input: dataset $\mathcal{D}$, randomly initialized RIN model R with feature extractor f_Φ and classifier g_Θ, randomly initialized SIN model S with feature extractor f_ϕ and classifier g_ω, randomly initialized and fixed teacher RIN model R_t with feature extractor $f_{t,\Phi}$ and classifier $g_{t,\Theta}$, epochs E, softmax operator σ, stopgrad operator SG, cross-entropy loss CE, Kullback-Leibler divergence loss KLD, mean square error MSE, Sobel edge operator h, feature alignment loss factors (γ_r, γ_r) , decision alignment loss factors (λ_r, λ_s), teacher-student decision alignment loss factor β

for epoch $e \in \{1, 2, .., E\}$ **do**
    sample a mini-batch $(x, y) \sim \mathcal{D}$
    $x_{shape} = h(x)$
    $z_\Phi = f_\Phi(x)$
    $z_\phi = f_\phi(x_{shape})$
    $R(x) = g_\Theta(f_\Phi(x))$
    $S(h(x)) = g_\omega(f_\phi(h(x_{shape})))$
    $R_t(x) = g_{t,\Theta}(f_{t,\Phi}(x))$
    $\mathcal{L}_{CER} = CE(\sigma(R(x)), y)$
    $\mathcal{L}_{CES} = CE(\sigma(S(h(x))), y)$
    $\mathcal{L}_{FAR} = MSE(z_\Phi, SG(z_\phi))$ ▷ (Eq. 5)
    $\mathcal{L}_{FAS} = MSE(SG(z_\Phi), z_\phi)$ ▷ (Eq. 6)
    $\mathcal{L}_{\mathcal{FA}} = \gamma_r\mathcal{L}_{FAR} + \gamma_s\mathcal{L}_{FAS}$ ▷ (Eq. 7)
    $\mathcal{L}_{DAR} = KLD(\sigma(R(x))), SG(\sigma(S(h(x))))$ ▷ (Eq. 8)
    $\mathcal{L}_{DAS} = KLD(SG(\sigma(R(x))), \sigma(S(h(x))))$ ▷ (Eq. 9)
    $\mathcal{L}_{\mathcal{DA}} = \lambda_r\mathcal{L}_{DAR} + \lambda_s\mathcal{L}_{DAS}$ ▷ (Eq. 10)
    $\mathcal{L}_{TS} = KLD(\sigma(R(x)), \sigma(R_t(x)))$
    $\mathcal{L} = \mathcal{L}_{CER} + \mathcal{L}_{CES} + \mathcal{L}_{FA} + \mathcal{L}_{DA} + \beta\mathcal{L}_{TS}$
    Update parameters of R and S based on $\mathcal{L}$ using Stochastic Gradient Descent
    Update R_t as EMA of R
end for
return RIN student model R

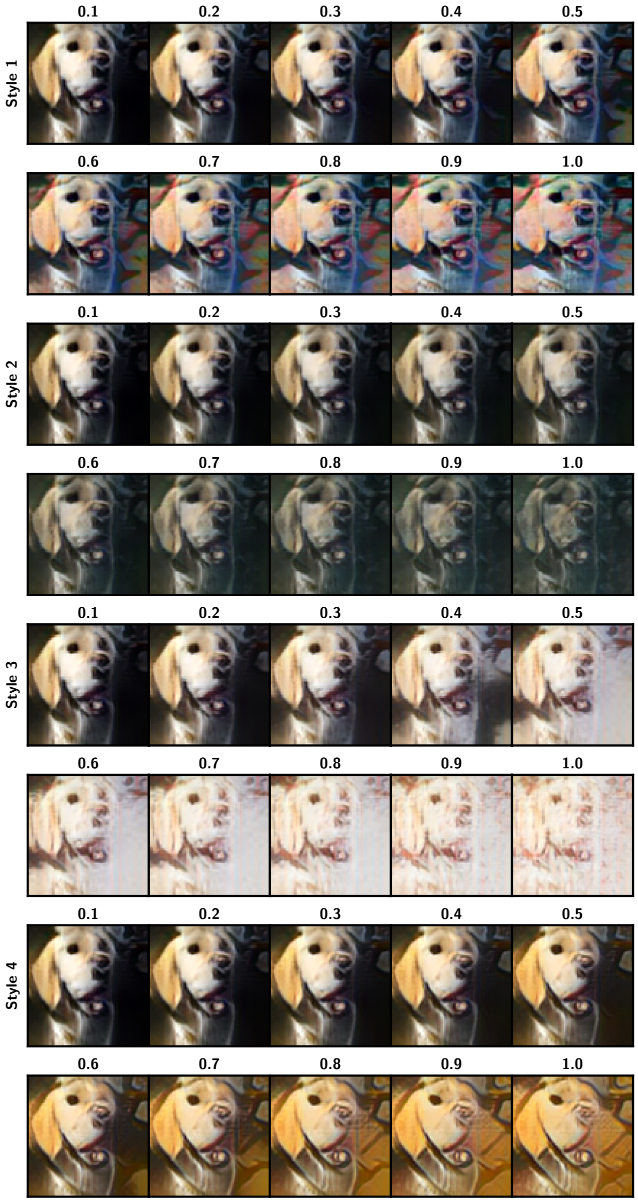

Figure 5. An illustration of the various stylized miniImageNet images generated for varying stylization intensities to perform the texture bias analysis in Section 6.1.

Figure 6. An illustration of class-specific tinted images generated for spurious correlation analysis is provided in Section 6.2.

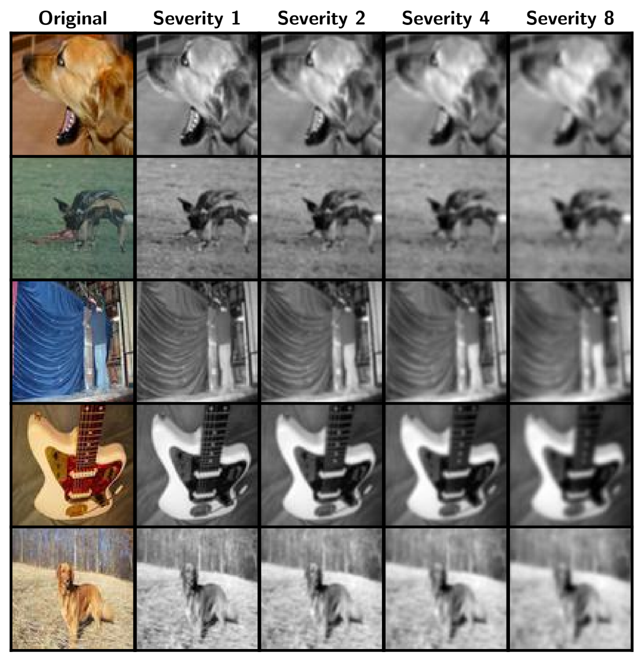

Figure 7. An illustration of low-pass filtered images generated for statistical regularity analysis in Section 6.3.