the last term on the right hand for optimal loss drop, while there would be no such restriction if the term was on the left hand. Therefore, if there is a Lipschitz constant for opposite direction denoted by $M > 0$ such that $||g_w(x_1) - g_w(x_2)|| \geq M||x_1 - x_2|| \forall x_1, x_2$), the following theorem can be derived.

Theorem 2. (The second theorem on gradient matching optimization). Suppose $||g_w(x_1) - g_w(x_2)|| \geq M||x_1 - x_2|| \forall x_1, x_2$ for $M > 0$ holds and $\mathcal{L}_{grad}(x) = ||g_w(x) - g^*||_2^2$ is the gradient matching loss. Then, when gradient descent Δx is applied with step size $\mu < \delta_1$ for some $\delta_1 > 0$, the following holds:

$$\mathcal{L}_{grad}(x + \mu\Delta x) \geq \mathcal{L}_{grad}(x) - \frac{1}{4M^2}||\frac{\partial \mathcal{L}_{grad}(x)}{\partial x}||_2^2, \quad (3)$$

where $\mu < \delta_1$ satisfies $||\mu\Delta x|| < \delta$ such that $g_w(x + \mu\Delta x) = g_w(x) + \mu\nabla_x g_w(x)^T\Delta x$ holds approximately.

Proof. See Appendix A2. □

The proof of Theorem is similar to that of Theorem except that the term including μ to be minimized is on the left side, thus there is no restriction like $\mu = \frac{1}{2L^2}$ in Theorem , thus more favorable. Inequality (3) implies that the upper bound of gradient matching loss drop is $\frac{1}{4M^2}||\frac{\partial \mathcal{L}_{grad}(x)}{\partial x}||_2^2$, unless the gradient term $\frac{\partial \mathcal{L}_{grad}(x)}{\partial x}$ is zero. Then, a gradient descent with a large M can hinder the convergence of gradient matching optimization. Therefore, we hypothesize that a global model with smaller M has a potential to experience a sharper loss drop in gradient matching optimization.

Finding L and M near Ground Truth: Maximum and Minimum Eigenvalues of Hessian

Theorems and are about one-step loss drop, so summarizing the whole process of optimization with them is difficult. To mitigate such complexity, we consider the loss drop near ground truth as it is the most important to decide whether ground truth can be reached through optimization or not. For a neighborhood point of of ground truth x^*, $x^* + \Delta x$ ($||\Delta x||$ is very small), gradient matching loss can be approximated by Taylor's second-order approximation like the following:

$$\mathcal{L}_{grad}(x^* + \Delta x) = \mathcal{L}_{grad}(x^*) + \nabla_{x=x^*}\mathcal{L}_{grad}(x)^T\Delta x + \frac{1}{2}\Delta x^T H(x^*)\Delta x,$$

where $H(x^*)$ is the Hessian of gradient matching loss with respect to input variables at x^*. Note that both $\mathcal{L}_{grad}(x^*)$ and $\nabla_{x=x^*}\mathcal{L}_{grad}(x)$ are zero when $\mathcal{L}_{grad}$ is either L2 or cosine distance. Then, $\mathcal{L}_{grad}(x^* + \Delta x)$ can be interpreted as the distance in gradient space while $||\Delta x||^2$ corresponds to distance in input space. Combining two preceding observations, the ratio of gradient distance to input distance is $\mathcal{L}_{grad}(x^* + \Delta x)/||\Delta x||^2 = \frac{1}{2}\frac{\Delta x}{||\Delta x||}^T H(x^*)\frac{\Delta x}{||\Delta x||}$. Then, the upper and lower bounds of this ratio correspond to maximum and minimum eigenvalues of Hessian, respectively. In proximity to ground turth, we can replace L and M with maximum and minimum eigenvalues of Hessian at ground truth, which is our proposed proxy, LAVP.

Hessian of Gradient Matching Loss

To compute LAVP, Hessian should be identified first. In Theorems 4 and 5, we derive the Hessian for L2 and cosine distances in a closed form respectively.

Theorem 3. (Hessian at ground truth for L2 distance). Suppose $\mathcal{L}_{grad}$ is L2 distance, then the Hessian at ground truth is like the following:

$$H_{\text{L2}}(x^*) = J(x^*)^T J(x^*), \quad (4)$$

where $J(x^*) = \nabla_{x=x^*} g_w(x)$ is the Jacobian of gradient function $g_w(x)$ with respect to input at ground truth x^*.

Proof. See Appendix A3. □

When $\mathcal{L}_{grad}$ is L2 distance, $H_{\text{L2}}(x^*) = J(x^*)^T J(^*)$ holds by Theorem , thus positive semi-definite. since input dimension is smaller than weight dimension in general, $H_{\text{L2}}(x^*)$ is not trivial low rank and its minimum eigenvalue has a potential to be positive.

For cosine distance, Hessian at ground truth can be solved in closed form by the following theorem.

Theorem 4. (Hessian at ground truth for cosine distance). Suppose $\mathcal{L}_{grad}$ is cosine distance, then the Hessian at ground truth is like the following:

$$H_{\cos}(x^*) = \frac{1}{||g^*||^2} J(x^*)^T (I - \frac{g^*}{||g^*||}\frac{g^{*T}}{||g^*||}) J(x^*), \quad (5)$$

where I is the identity matrix.

Proof. See Appendix A4. □

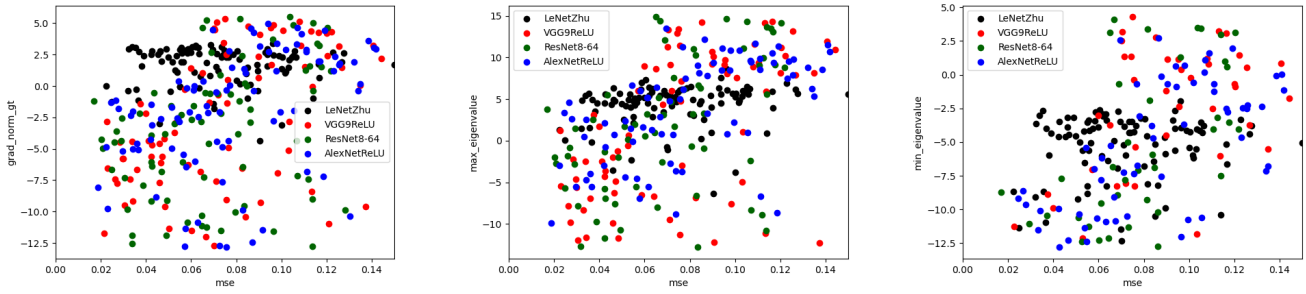
The minimum eigenvalue of $H_{\cos}(x^*)$ is nonnegative as it is positive semi-definite by Cauchy-Schwartz inequality.

Implementation of LAVP

To find the maximum eigenvalue of Hessian, power iteration is used. Power iteration computes matrix-vector product and normalization alternatively until the vector converges to the eigenvector with the maximum eigenvalue. When this algorithm is applied to the Hessian, Jacobian-vector product is inevitable, while *Autograd* package in PyTorch supports only vector-Jacobian product. Therefore, Jacobian-vector product is solved with the finite difference method with very small step size. Once the maximum eigenvalue α_{max} is obtained for the Hessian $H(x^*)$, then power iteration is applied to $\alpha_{max}I - H(x^*)$ (I is the identity matrix) for identifying the minimum eigenvalue α_{min}, as $\alpha_{max} - \alpha_{min}$ would be the maximum eigenvalue of $\alpha_{max}I - H(x^*)$ (I is the identity matrix). *It is noteworthy that multiple Hessian-vector products, rather than the entire Hessian, are sufficient for computing LAVP, thus more efficient.*

σ_S	grad_norm	max (LAVP for L2)	min (LAVP for L2)	ang_max (LAVP for CS)	ang_min (LAVP for CS)
C-10+L2	0.35 / -0.27 / -0.35 / -0.13	**0.51** / **-0.46** / **-0.51** / -0.15	0.41 / -0.40 / -0.41 / **-0.20**	-0.06 / -0.01 / 0.06 / -0.06	-0.04 / -0.07 / 0.04 / -0.06
C-100+L2	0.41 / -0.31 / -0.41 / -0.19	**0.46** / **-0.57** / **-0.46** / 0.10	0.41 / -0.45 / -0.41 / -0.01	-0.05 / -0.18 / 0.05 / **0.22**	-0.05 / -0.20 / 0.05 / 0.19
IN+L2	0.03 / 0.03 / -0.03 / 0.19	0.33 / -0.26 / -0.33 / **0.49**	**0.34** / -0.25 / **-0.35** / 0.46	0.14 / -0.20 / -0.14 / 0.42	0.25 / **-0.34** / -0.25 / 0.42
IW+L2	0.35 / -0.01 / -0.35 / 0.00	0.66 / **-0.58** / -0.66 / 0.46	**0.68** / -0.52 / **-0.68** / 0.29	0.38 / -0.41 / -0.38 / 0.46	0.46 / -0.52 / -0.46 / **0.49**
C-10+CS	-0.28 / 0.25 / 0.28 / -0.18	-0.03 / -0.02 / 0.03 / 0.00	0.04 / -0.08 / -0.04 / 0.02	0.26 / -0.31 / -0.26 / 0.36	**0.64** / **-0.74** / **-0.64** / **0.62**
C-100+CS	0.10 / -0.07 / -0.10 / 0.02	0.37 / -0.35 / -0.37 / 0.26	0.32 / -0.34 / -0.32 / 0.24	0.67 / -0.8 / -0.67 / **0.68**	**0.69** / **-0.81** / **-0.69** / 0.65
IN+CS	-0.13 / 0.11 / 0.14 / -0.18	0.16 / -0.16 / -0.16 / 0.12	0.27 / -0.28 / -0.25 / 0.17	0.57 / -0.65 / -0.57 / 0.64	**0.75** / **-0.80** / **-0.75** / **0.73**
IW+CS	0.00 / 0.01 / 0.00 / -0.04	0.37 / -0.37 / -0.37 / 0.21	0.33 / -0.34 / -0.33 / 0.18	0.72 / -0.73 / -0.72 / **0.61**	**0.75** / **-0.83** / **-0.75** / 0.61

Table 1: **Spearman's correlation (σ_S) of proxy candidates with image similarity scores (MSE(↓) / SSIM(↑) / PSNR(↑) / LPIPS(↓)) on low-resolution images.** 'C-10', 'C-100', 'IN', and 'IW' denote CIFAR-10, CIFAR-100, ImageNette, and ImageWoof respectively. 'L2' and 'CS' denote L2 and cosine distances respectively. 'grad_norm', 'max', 'min', 'ang_max', and 'ang_min' denote the gradient norm, the maximum eigenvalue of Hessian for L2, the minimum eigenvalue of Hessian for L2, the maximum eigenvalue of Hessian for CS, and the minimum eigenvalue of Hessian for CS.



(a) grad_norm vs MSE (L2, $\sigma_S = 0.35$) (b) max vs MSE (L2, $\sigma_S = \mathbf{0.51}$) (c) min vs MSE (L2, $\sigma_S = 0.41$)

(d) grad_norm vs MSE (CS, $\sigma_S = -0.28$) (e) ang_max vs MSE (CS, $\sigma_S = 0.26$) (f) ang_min vs MSE (CS, $\sigma_S = \mathbf{0.64}$))

Figure 2: **Comparison of the gradient norm, maximum and minimum eigenvalues of Hessian in terms of the correlation with MSE of reconstructed samples over several architectures on CIFAR10 test samples.**

Experimental Results

In this section, we begin with a concise overview of our experimental setup. Then, we elucidate the advantages of LAVP over the gradient norm (baseline), in explaining the vulnerability to the gradient inversion attack with either L2 or cosine distance by providing correlation tables and plots. For a black-box scenario where the attacker's loss function is unknown to clients, we also introduce the loss-agnostic LAVP fusion, the proxy that can handle several candidate loss functions at once. An example of LAVP fusion includes the geometric mean between LAVPs for L2 and cosine distances. We provide the correlation table for this example with a comparative evaluation against the gradient norm.

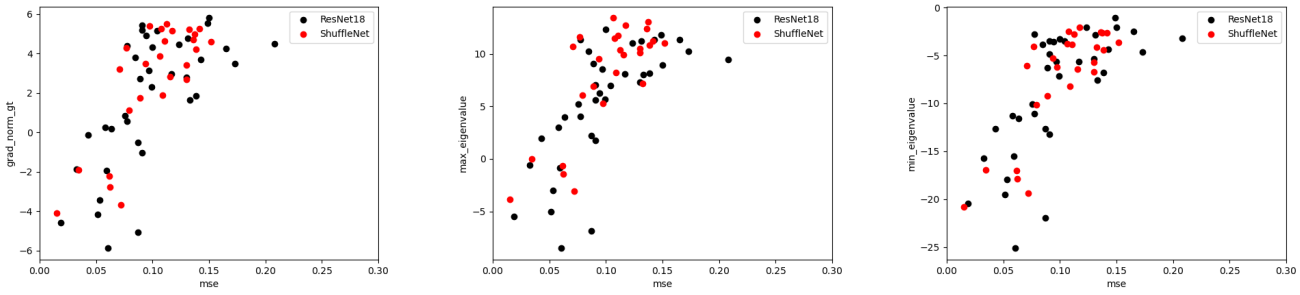
Experimental Setup

Datasets. We conducted an evaluation by randomly selecting 100 validation images from CIFAR-10 (Krizhevsky 2009), CIFAR-100 (Krizhevsky 2009), ImageNette (Howard), ImageWoof (Howard), and ImageNet (Deng et al. 2009). Notably, ImageNette and ImageWoof are subsets of ImageNet (Deng et al. 2009), each consisting of ten easily classified classes, but with different classes from one another.

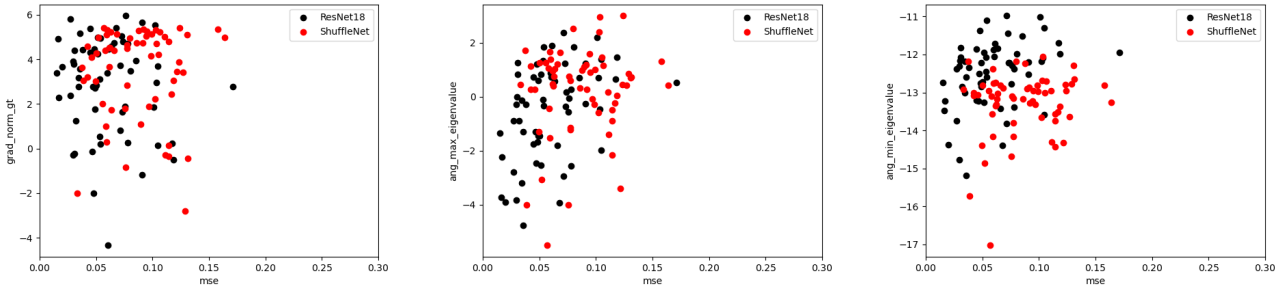
Architectures and Attack Hyperparameters. We evaluated several deep learning models on low-resolution images, including LeNet (LeCun et al. 1998), AlexNet (Krizhevsky, Sutskever, and Hinton 2017), VGG9 (Simonyan and Zisserman 2014), and ResNet8 (He et al. 2016). We trained these models on a training set for 300 epochs, using the SGD optimizer with an initial learning rate of 0.1 and a learning rate

σ_S	grad_norm	max (LAVP for L2)	min (LAVP for L2)	ang_max (LAVP for CS)	ang_min (LAVP for CS)
ImageNet+L2	0.66 / -0.66 / -0.66 / **-0.28**	0.69 / -0.72 / -0.69 / -0.09	**0.74** / **-0.74** / **-0.72** / -0.07	-0.05 / 0.09 / 0.05 / 0.03	-0.07 / 0.12 / 0.07 / 0.10
ImageNet+CS	-0.06 / -0.04 / 0.00 / -0.05	0.02 / -0.07 / -0.11 / 0.04	-0.03 / -0.05 / -0.06 / 0.00	**0.27** / **-0.37** / -0.21 / 0.32	0.26 / -0.22 / **-0.24** / **0.32**

Table 2: **Spearman's correlation of proxy candidates with image similarity scores (MSE(↓) / SSIM(↑) / PSNR(↑) / LPIPS(↓)) on ImageNet.**



(a) grad_norm vs MSE (L2, $\sigma_S = 0.66$) (b) max vs MSE (L2, $\sigma_S = 0.69$) (c) min vs MSE (L2, $\sigma_\mathbf{S} = \mathbf{0.74}$)



(d) grad_norm vs MSE (CS, $\sigma_S = -0.06$) (e) ang_max vs MSE (CS, $\sigma_\mathbf{S} = \mathbf{0.27}$) (f) ang_min vs MSE (CS, $\sigma_S = 0.26$))

Figure 3: **Comparison of the gradient norm, maximum and minimum eigenvalues of Hessian in terms of the correlation with MSE of reconstructed samples over ResNet18 and ShuffleNet models on ImageNet validation samples.**

σ_S	grad_norm	$\sqrt{\text{max} * \text{ang\_min}}$ (LAVP fusion)
C-10+L2	0.35 / -0.27 / -0.35 / **-0.13**	**0.48** / **-0.49** / **-0.48** / -0.09
C-100+L2	0.41 / -0.31 / -0.41 / **-0.19**	**0.50** / **-0.49** / **-0.50** / -0.09
IN+L2	0.03 / 0.03 / -0.03 / 0.19	**0.48** / **-0.41** / **-0.48** / **0.59**
IW+L2	0.35 / -0.01 / -0.35 / 0.00	**0.71** / **-0.71** / **-0.71** / **0.51**
C-10+CS	-0.21 / **0.25** / -0.21 / 0.15	**-0.28** / **0.25** / **0.28** / **-0.18**
C-100+CS	0.10 / -0.07 / -0.10 / 0.02	**0.50** / **-0.45** / **-0.50** / **0.36**
IN+CS	-0.13 / 0.11 / 0.14 / -0.18	**0.44** / **-0.43** / **-0.44** / **0.34**
IW+CS	0.00 / 0.01 / 0.00 / -0.04	**0.57** / **-0.59** / **-0.57** / **0.38**

Table 3: **Spearman's correlation of loss-agnostic LAVP and gradient norm with image similarity scores (MSE(↓) / SSIM(↑) / PSNR(↑) / LPIPS(↓)).** An instance of LAVP fusion is the geometric mean between the maximum eigenvalue of Hessian for L2 distance and the minimum eigenvalue of Hessian for cosine distance.

decay of 0.1 at the 150^{th} and 225^{th} epochs. We also trained ResNet18 (He et al. 2016) and ShuffleNet (Ma et al. 2018) models on high-resolution images from ImageNet. To assess the vulnerability to attacks, we directly performed gradient inversion attacks on 100 validation images randomly selected from each dataset. We used attack algorithms from previous works (Geiping et al. 2020; Yin et al. 2021; Zhu, Liu, and Han 2019) and considered two major gradient matching losses: L2 and cosine distances. Also, we incorporated the total variation loss for regularization. We use Adam optimizer (Kingma and Ba 2015) for gradient inversion. For each sample, we run attack algorithm three times using different random seeds. The final outcome was recorded the one that reconstructs the best among these runs.

Image Similarity Scores. Image similarity scores measure the quality of reconstructed images compared to the original images. We consider Mean Squared Error (MSE), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), Structural Similarity Index (SSIM) (Wang et al. 2004) and Peak Signal-to-Noise Ratio (PSNR) for quantifying reconstruction quality. MSE computes the mean pixel-wise difference between original sample and its reconstruction in image space. LPIPS computes the distance from ground truth within the feature space of the ImageNet-pretrained VGG network. SSIM measures the similarity between two images by comparing their structural information, luminance, and contrast. PSNR measures the quality of reconstructed images using signal-to-noise (SNR) ratio.

Proxy for the Vulnerability. We consider the gradient norm and LAVP as the candidates for the proxy. In tables, the gradient norm is denoted as 'grad_norm', maximum and minimum eigenvalues for L2 distance hessian are denoted as 'max' and 'min', and maximum and minimum eigenvalues for cosine distance loss are denoted as 'ang_max' and 'ang_min'. Here, 'ang' is the abbreviation for 'angular'.