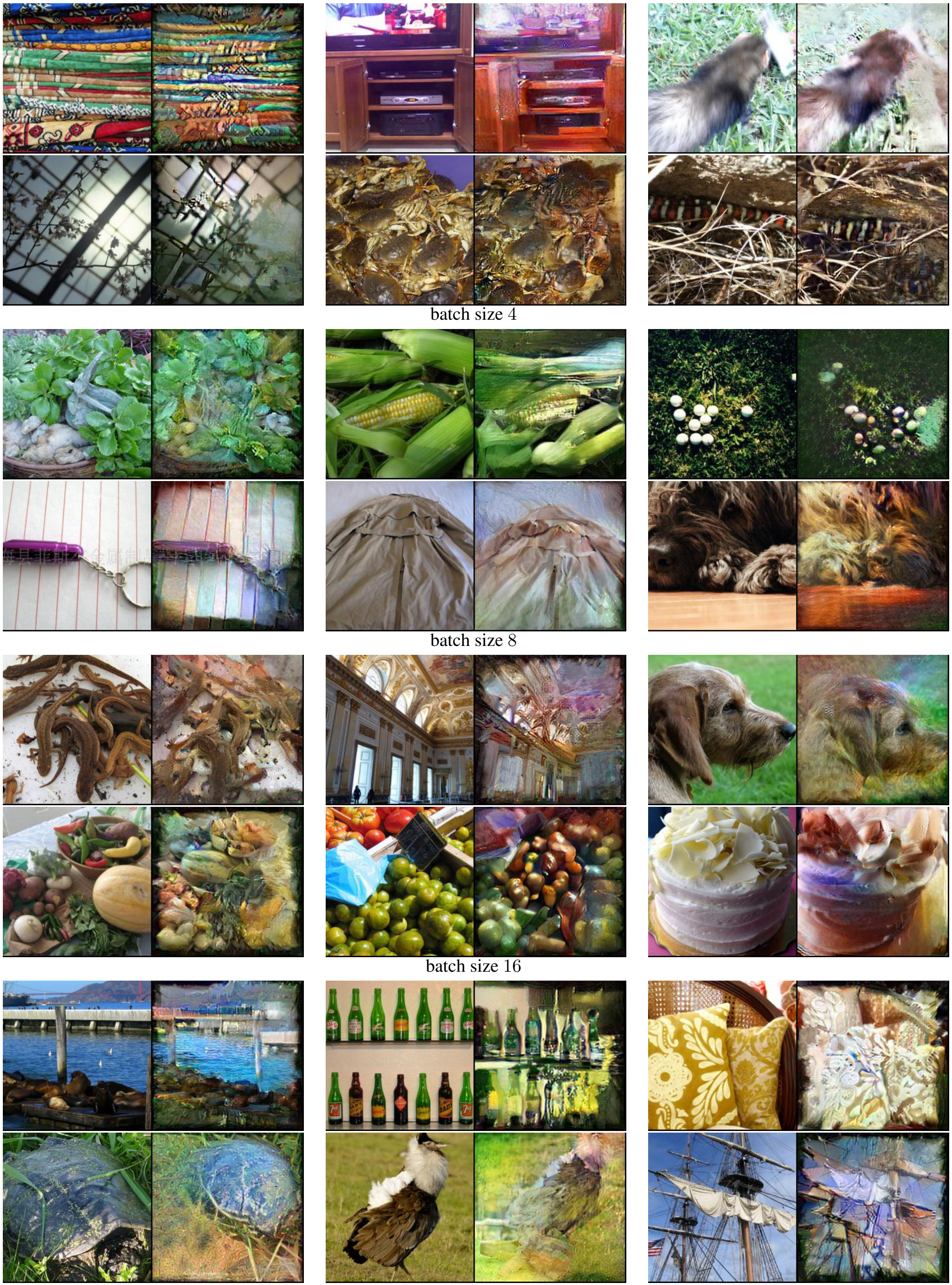
Appendix A - More Examples



batch size 4

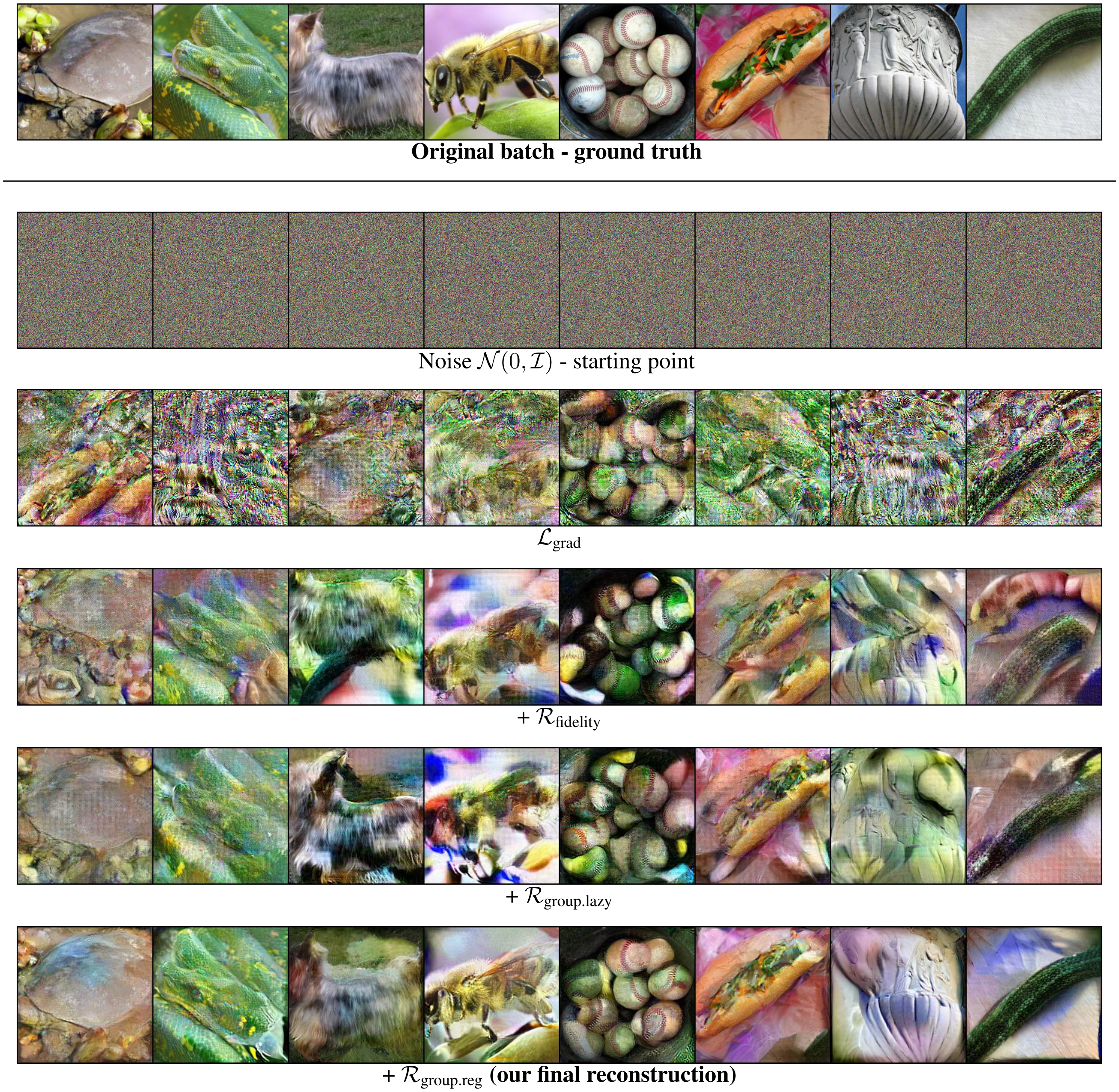
batch size 8

batch size 16

batch size 32

Figure 9: Additional examples of information leakage when inverting ResNet-50 gradients on the ImageNet validation set. Each block containing a pair of (left) original sample and its (right) reconstruction by GradInversion.

Appendix B - Ablation Studies Images



Original batch - ground truth

Noise $\mathcal{N}(0, \mathcal{I})$ - starting point

$\mathcal{L}_{\text{grad}}$

+ $\mathcal{R}_{\text{fidelity}}$

+ $\mathcal{R}_{\text{group.lazy}}$

+ $\mathcal{R}_{\text{group.reg}}$ **(our final reconstruction)**

Figure 10: Detailed visual comparison when adding individual loss terms to GradInversion (supplementary for Table 5 of main paper). Starting from noise, the gradient loss produces noisy outputs which begin to show glimpses of what the original image contains. The fidelity loss encourages the optimization to produce more realistic outputs. Using multiple random seeds and regularizing the inputs using even the simple lazy scheme of conforming to the mean image improves the image quality. Our group-based regularization that uses image registration produces the best-looking outputs.

Appendix C - Additional Details & Analysis

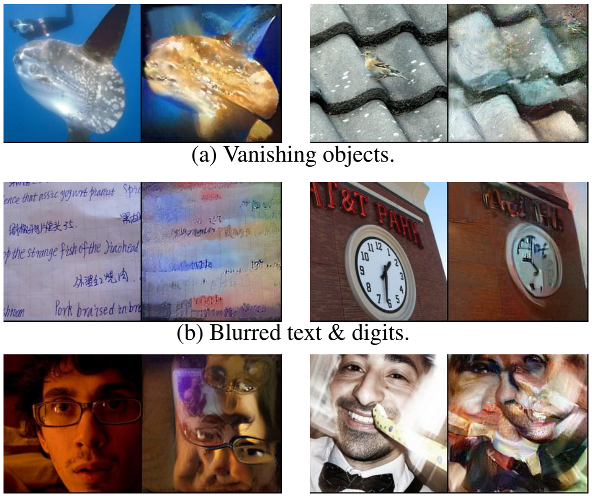
$\mathcal{L}_{grad}$ cost function. For the task of gradient matching, we study how the loss function affects optimization in Table 5. We find ℓ_2 loss outperforms cosine similarity [13] for gradient matching. To this end, we compare the ℓ_2 distance, percentage of gradient signs that matched, and the cosine similarity between the final optimized gradient and the ground truth gradient. It can be observed that the ℓ_2 loss results in stronger convergence for this task.

$\mathcal{L}_{\text{grad}}(\cdot)$	$\nabla_{\mathbf{W}}\mathcal{L}(\hat{\mathbf{x}}^*, \hat{\mathbf{y}}^*)$ **vs.** $\nabla_{\mathbf{W}}\mathcal{L}(\mathbf{x}^*, \mathbf{y}^*)$		
	ℓ_2 dist. ↓	sign (%) ↑	cos. dist. ↓
cosine [13]	5.965	79.0	0.139
ℓ_2 (ours)	**3.835**	**80.9**	**0.110**

Table 5: Comparing the efficacy for cosine and ℓ_2 distance for gradient matching.

Insights & open challenges. We next discuss several of our observations when performing GradInversion for the ResNet-50 network (MOCO V2) on the ImageNet1K dataset. We would like to note that these observations hold for the chosen network and optimization settings used, and may not be general in scope. We hope that sharing our experiences would help provide insights for future work.

- **Vanishing objects.** Images recovered from gradient inversion occasionally omit details of original images, as shown in Fig. 11 (a) where the diver and the bird disappear post inversion. The observation is in line with the missing details phenomena during the latent code projection as in StyleGAN2 by Karras *et al*. [23].
- **Texts & digits.** GradInversion unveils existence of texts and digits, while their exact details remain blurry. See several quick examples in Fig. 11 (b).
- **Human faces.** Recovery remains harder for samples and distributions in ImageNet that involve human faces (also deemed challenging by BigGAN [4]). Even though detailed features and patches can be reversed, *e.g*., mouths, eyes, and noses, *etc*., they are not correctly arranged spatially, as shown in Fig. 11 (c). We conjecture that this is a result of (i) the under-representation of such distributions in ImageNet1K and (ii) such features being ignored by the network for the classification task. Different observations may hold for other datasets where tasks enforce networks to focus on spatial alignments of facial features, *e.g*., towards facial recognition.



(a) Vanishing objects.

(b) Blurred text & digits.

(c) Human faces.

Figure 11: Insights and observations of GradInversion given ResNet-50 gradients on ImageNet. Each block containing a pair of (left) original sample and its (right) reconstruction by GradInversion. Samples from inversion results at batch sizes 4 and 8.