Appendix for *Dense Vision Transformer Compression with Few Samples*

In this appendix, we present additional details and results to complement the main paper. We first provide the full results of the latency analysis on the 12 blocks of the ViT-Base model. Then we present visualizations of synthetic metric sets produced by various ViT model variants, and the time for updating synthetic images.

A. Latency analysis

We conducted latency tests on the 12 blocks of the ViT-Base model, analyzing the impact of completely or partially removing each block. The term "Attn+0.5·FFN" denotes the standard approach of DC-ViT in compressing a ViT block, which involves eliminating the entire attention module and a portion of the MLP module. We tested the latency 5 times with 5 random seeds (2021 $\sim$ 2025) and reported the mean latency with the standard deviation. The full results are detailed in Tab. 8. Given the structural uniformity of all ViT blocks within the same model, our findings indicate minimal latency variance when different blocks are compressed using the DC-ViT approach.

Latency (ms)	Block	Attn
0	$102.29_{\pm 0.19}$	$107.50_{\pm 0.26}$
1	$102.85_{\pm 0.03}$	$107.64_{\pm 0.14}$
2	$102.81_{\pm 0.19}$	$107.69_{\pm 0.08}$
3	$102.89_{\pm 0.14}$	$107.71_{\pm 0.08}$
4	$102.88_{\pm 0.23}$	$107.53_{\pm 0.14}$
5	$103.00_{\pm 0.06}$	$107.66_{\pm 0.10}$
6	$102.97_{\pm 0.12}$	$107.74_{\pm 0.07}$
7	$102.89_{\pm 0.12}$	$107.73_{\pm 0.06}$
8	$102.82_{\pm 0.16}$	$107.87_{\pm 0.07}$
9	$102.84_{\pm 0.05}$	$107.80_{\pm 0.19}$
10	$102.96_{\pm 0.18}$	$107.82_{\pm 0.15}$
11	$102.84_{\pm 0.23}$	$107.68_{\pm 0.10}$
Latency (ms)	FFN	Attn+0.5·FFN
0	$110.99_{\pm 0.36}$	$103.84_{\pm 0.22}$
1	$111.18_{\pm 0.06}$	$103.97_{\pm 0.18}$
2	$111.13_{\pm 0.12}$	$104.04_{\pm 0.09}$
3	$111.14_{\pm 0.23}$	$104.03_{\pm 0.08}$
4	$111.15_{\pm 0.07}$	$104.00_{\pm 0.13}$
5	$111.23_{\pm 0.06}$	$104.09_{\pm 0.09}$
6	$111.06_{\pm 0.23}$	$103.98_{\pm 0.06}$
7	$111.10_{\pm 0.11}$	$103.92_{\pm 0.10}$
8	$111.02_{\pm 0.20}$	$103.74_{\pm 0.40}$
9	$111.07_{\pm 0.16}$	$103.72_{\pm 0.51}$
10	$111.20_{\pm 0.06}$	$103.88_{\pm 0.19}$
11	$111.19_{\pm 0.02}$	$103.89_{\pm 0.28}$

Table 8. The latency in milliseconds (ms) of ViT-Base when different parts are removed across the 12 blocks. "Attn": the attention module, "FFN": the feed-forward network (MLP), and "Attn+0.5·FFN": the attention module plus half of MLP.

B. Synthetic Metric Sets

We present visualizations of synthetic metric sets produced by various ViT model variants, as shown from Fig. 5 to Fig. 8. It's evident that these synthetic images successfully capture the distribution of the original dataset. Initially, the images start from Gaussian noise with randomly assigned class labels, yet they evolve to showcase substantial semantic content, capturing object textures, shapes, and complex details, which is a testament to the effectiveness of our generation process.

On the other hand, even though the images we generate are still far from being as realistic as actual samples, the compressed model's metric loss on synthetic data can still reflect its true performance on the test set very accurately. This balance between simplicity in generation and accuracy in performance assessment marks a significant stride in using synthetic data for model evaluation.

As shown in Tab. 9, we also tested the latency of one iteration for different ViT variants to update synthetic images. The time is measured on a single RTX 3090 GPU. Compared to the training time, we can see that for different models, the total time taken to generate synthetic data is very short, amounting to less than 10% of the duration of a single finetune after compressing one block.

	ViT-T	ViT-S	ViT-B	ViT-L	DeiT-B	Swin-B
Latency (ms)	296.2	526.5	985.5	3204.2	302.4	390.1

Table 9. The time of one iteration for different ViT variants to update synthetic images.

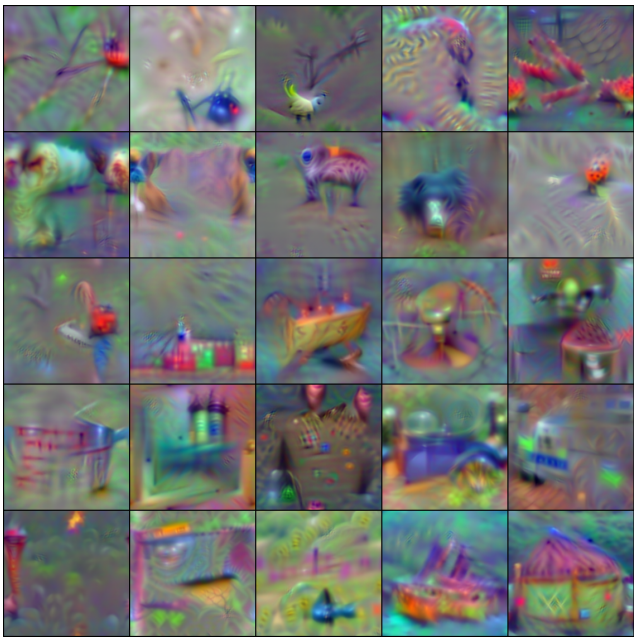

Figure 5. The synthetic metric set generated by ViT-Tiny.

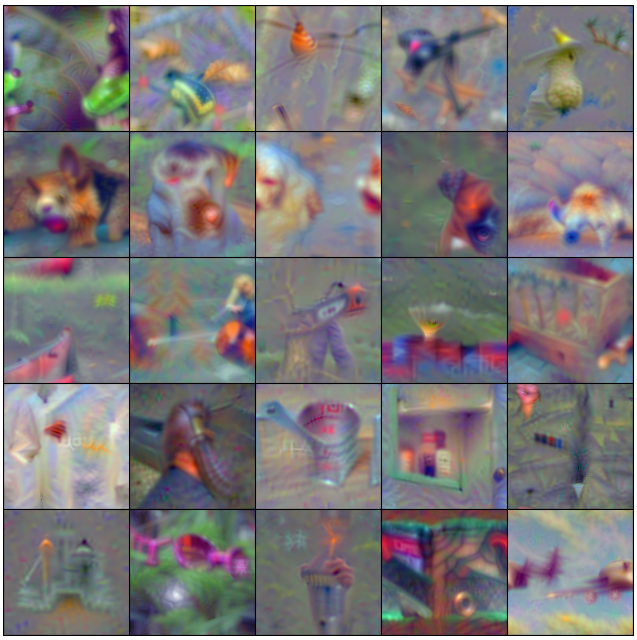

Figure 7. The synthetic metric set generated by ViT-Base.

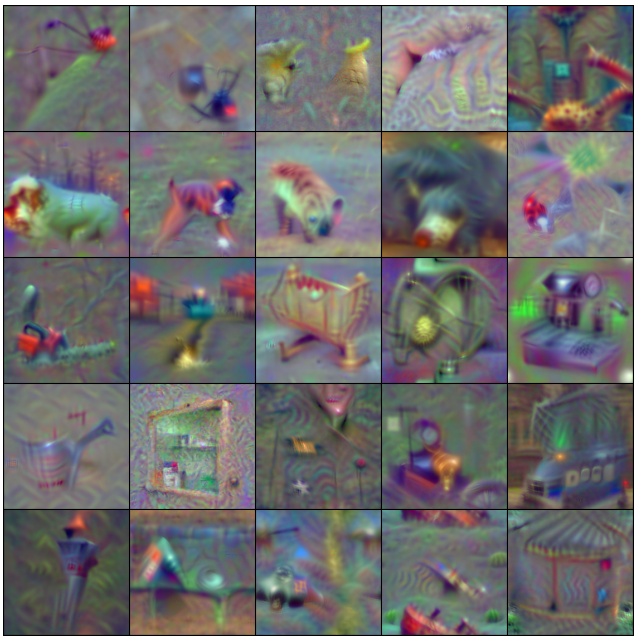

Figure 6. The synthetic metric set generated by ViT-Small.

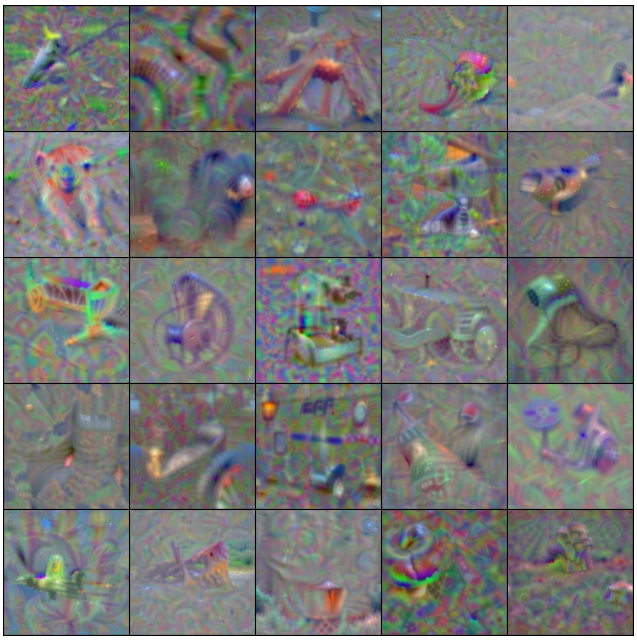

Figure 8. The synthetic metric set generated by ViT-Large.