Dataset Distillation with Neural Characteristic Function: A Minmax Perspective

Supplementary Material

A. Detailed Experimental Settings

A.1. Computational Resources

All experiments were conducted on NVIDIA GPUs, specifically the RTX 4090 and A100. For the experiments reported in Tables 1 and 2, all NCFM results were conducted on a single NVIDIA 4090 GPU. The comparative experiments between NCFM and DATM presented in Figure 3 were executed on a cluster of 8 A100 GPUs. The evaluation of computational efficiency discussed in Section 5.2 was conducted on a single A100 GPU.

A.2. Hyper-parameter Settings

We report the comprehensive hyper-parameter settings of our method in Table 7. The learning rate and weight decay are carefully tuned for different datasets and IPC settings to ensure optimal performance. We employ a fixed batch size of 1024 for all training processes, except for the ImageNet subset where we use a batch size of 512 when the IPC is ≥ 10. The amplitude-phase ratio α in the Neural Characteristic Function Discrepancy (NCFD) measure is set to 0.5, striking a balance between amplitude and phase components for optimal performance. This setting ensures equal contribution from both components, as demonstrated in our ablation study in Figure 5. The number of sampled frequency arguments $\boldsymbol{t}$ is consistently set to 4096 across all experiments. We set the factor of $\rho = 1$ when IPC > 50 and 2 otherwise.

Table 7. Hyper-parameter settings for different datasets and IPC configurations.

Dataset	IPC	Learning Rate	Weight Decay
CIFAR-10	1	0.0001	0.001
	10	0.001	0.001
	50	0.005	0.001
	500	5e-5	0.0001
	1000	0.0001	0.0001
CIFAR-100	1	0.0001	0.001
	10	0.001	0.001
	50	0.005	0.001
	100	1e-5	0.0001
Tiny ImageNet	1	0.0001	0.001
	10	0.001	0.001
	50	0.005	0.001
ImageNet subsets	1	0.0001	0.001
	10	0.001	0.001
	50	5e-5	0.0001

B. Proof of Theorem 2

Theorem 2 (Uniqueness for Characteristic Functions) *If two random variables X and Y have the same characteristic function, $\Phi_X(\boldsymbol{t}) = \Phi_Y(\boldsymbol{t})$ for all $\boldsymbol{t}$, then X and Y are identically distributed. In other words, a characteristic function uniquely determines the distribution.*

Proof 2 *We aim to prove the Uniqueness for Characteristic Functions(Theorem 2), which states that if two random variables have identical characteristic functions, then they are identically distributed. The characteristic function of a random variable X is defined as:*

$$\Phi_X(\boldsymbol{t}) = \mathbb{E}\left[e^{j\langle \boldsymbol{t}, X\rangle}\right]$$

Characteristic functions possess several important properties: they satisfy normalization where $\Phi_X(\mathbf{0}) = 1$, exhibit conjugate symmetry, are positive definite, and maintain uniform continuity across their domain.

Uniqueness via Fourier Inversion: *Assume X and Y are two random variables such that $\Phi_X(\boldsymbol{t}) = \Phi_Y(\boldsymbol{t})$ for all $\boldsymbol{t} \in \mathbb{R}^d$. By the **Fourier Inversion Theorem**, the probability distribution of a random variable can be uniquely determined by its characteristic function. Specifically, the inverse Fourier transform of the characteristic function yields the probability density function (if it exists) or the probability measure. Therefore, since $\Phi_X(\boldsymbol{t}) = \Phi_Y(\boldsymbol{t})$ for all $\boldsymbol{t}$, it follows that the probability distributions of X and Y are identical. Formally:*

$$F_X(\boldsymbol{x}) = F_Y(\boldsymbol{x}) \quad \textit{for all } \boldsymbol{x} \in \mathbb{R}^d$$

This means that for any Borel set $A \subseteq \mathbb{R}^d$:

$$\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$$

Since the characteristic functions of X and Y are identical and the Fourier Inversion Theorem ensures that characteristic functions uniquely determine probability distributions, we conclude that X and Y are identically distributed. This completes the proof of the Uniqueness for Characteristic Functions.

C. Proof of Theorem 3

Theorem 3 (CFD as a Distance Metric) *The CF discrepancy $\mathcal{C}_{\mathcal{T}}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$, serves as a distance metric between $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ when the support of $\mathcal{T}$ resides in Euclidean space. It satisfies the properties of non-negativity, symmetry, and the triangle inequality.*

Proof 3 *We aim to prove that the Characteristic Function Discrepancy (CFD) $\mathcal{C}_{\mathcal{T}}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ satisfies the properties of a valid distance metric: non-negativity, symmetry, and the triangle inequality.*

Non-negativity: *By the definition of CFD as*

$$\mathcal{C}_{\mathcal{T}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \int_{\boldsymbol{t}} \sqrt{\mathit{Chf}(\boldsymbol{t})}\, dF_{\mathcal{T}}(\boldsymbol{t}),$$

where $\mathit{Chf}(\boldsymbol{t}) = (\Phi_{\boldsymbol{x}}(\boldsymbol{t}) - \Phi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t}))(\overline{\Phi}_{\boldsymbol{x}}(\boldsymbol{t}) - \overline{\Phi}_{\tilde{\boldsymbol{x}}}(\boldsymbol{t}))$. Since $\mathit{Chf}(\boldsymbol{t})$ represents the squared magnitude of the difference between the characteristic functions, it is always non-negative:

$$\mathit{Chf}(\boldsymbol{t}) = |\Phi_{\boldsymbol{x}}(\boldsymbol{t}) - \Phi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t})|^2 \geq 0.$$

Therefore, the integral of a non-negative function is also non-negative:

$$\mathcal{C}_{\mathcal{T}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \geq 0.$$

Moreover, $\mathcal{C}_{\mathcal{T}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = 0$ if and only if $\mathit{Chf}(\boldsymbol{t}) = 0$ for all $\boldsymbol{t} \in \mathbb{R}^m$. This implies that $\Phi_{\boldsymbol{x}}(\boldsymbol{t}) = \Phi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t})$ for all $\boldsymbol{t}$, and by the Uniqueness Theorem of characteristic functions (Theorem 2), it follows that $\boldsymbol{x} =^d \tilde{\boldsymbol{x}}$.

Symmetry: *The CFD is inherently symmetric with respect to its arguments. Specifically, swapping $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ does not change the value of $\mathit{Chf}(\boldsymbol{t})$:*

$$\begin{aligned}\mathit{Chf}(\boldsymbol{t}) &= (\Phi_{\boldsymbol{x}}(\boldsymbol{t}) - \Phi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t}))(\overline{\Phi_{\boldsymbol{x}}(\boldsymbol{t})} - \overline{\Phi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t})}) \\ &= (\Phi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t}) - \Phi_{\boldsymbol{x}}(\boldsymbol{t}))(\overline{\Phi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t})} - \overline{\Phi_{\boldsymbol{x}}(\boldsymbol{t})}).\end{aligned}$$

Thus,

$$\mathcal{C}_{\mathcal{T}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \mathcal{C}_{\mathcal{T}}(\tilde{\boldsymbol{x}}, \boldsymbol{x}).$$

Triangle Inequality: *Consider three random variables $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$. We need to show that*

$$\mathcal{C}_{\mathcal{T}}(\boldsymbol{x}, \boldsymbol{z}) \leq \mathcal{C}_{\mathcal{T}}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{C}_{\mathcal{T}}(\boldsymbol{y}, \boldsymbol{z}).$$

Using the Minkowski inequality for integrals, we have:

$$\begin{aligned}\mathcal{C}_{\mathcal{T}}(\boldsymbol{x}, \boldsymbol{z}) &= \int_{\boldsymbol{t}} \sqrt{|\Phi_{\boldsymbol{x}}(\boldsymbol{t}) - \Phi_{\boldsymbol{z}}(\boldsymbol{t})|^2}\, dF_{\mathcal{T}}(\boldsymbol{t}) \\ &= \int_{\boldsymbol{t}} |\Phi_{\boldsymbol{x}}(\boldsymbol{t}) - \Phi_{\boldsymbol{z}}(\boldsymbol{t})|\, dF_{\mathcal{T}}(\boldsymbol{t}) \\ &= \int_{\boldsymbol{t}} |\Phi_{\boldsymbol{x}}(\boldsymbol{t}) - \Phi_{\boldsymbol{y}}(\boldsymbol{t}) + \Phi_{\boldsymbol{y}}(\boldsymbol{t}) - \Phi_{\boldsymbol{z}}(\boldsymbol{t})|\, dF_{\mathcal{T}}(\boldsymbol{t}) \\ &\leq \int_{\boldsymbol{t}} (|\Phi_{\boldsymbol{x}}(\boldsymbol{t}) - \Phi_{\boldsymbol{y}}(\boldsymbol{t})| + |\Phi_{\boldsymbol{y}}(\boldsymbol{t}) - \Phi_{\boldsymbol{z}}(\boldsymbol{t})|)\, dF_{\mathcal{T}}(\boldsymbol{t}) \\ &= \mathcal{C}_{\mathcal{T}}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{C}_{\mathcal{T}}(\boldsymbol{y}, \boldsymbol{z}).\end{aligned}$$

Thus, the triangle inequality holds.

Since $\mathcal{C}_{\mathcal{T}}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ satisfies non-negativity, symmetry, and the triangle inequality, it qualifies as a valid distance metric between the random variables $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$. This completes the proof of the CFD as a Distance Metric.

D. Results on Larger IPC Datasets

To comprehensively validate the scalability of NCFM, we conducted experiments on datasets with larger IPC settings.

CIFAR-10/100. As shown in Table 8, NCFM demonstrates superior performance across different IPC configurations. Notably, several traditional methods, including MTT, TESLA, and FTD, fail to match even the random baseline in high IPC settings, while NCFM maintains robust performance. For CIFAR-100 with 100 IPC, NCFM achieves 58.8%, outperforming DATM by 1.3% and even exceeding the whole dataset performance. This efficient performance demonstrates that NCFM can achieve lossless dataset distillation with significantly reduced memory requirements, using about merely 2GB GPU memory.

Higher-resolution Datasets. Additionally, we examined NCFM's effectiveness on ImageNet subsets with larger IPC settings, as presented in Table 9. Our method demonstrates consistent improvements across all subsets. Most remarkably, **NCFM achieves lossless dataset distillation on multiple ImageNet subsets, even surpassing the whole dataset performance** in most cases. Compared to the current SOTA method RDED, NCFM achieves significant improvements of 7.7% and 3.0% on ImageWoof and ImageNette respectively. For instance, on ImageFruit, our method outperforms the whole dataset training by 5.1%, demonstrating its exceptional capability in condensing high-resolution datasets.

Table 8. Performance comparison (%) on CIFAR-10 and CIFAR-100 with larger IPC settings. "↓ random" indicates performance below random selection baseline.

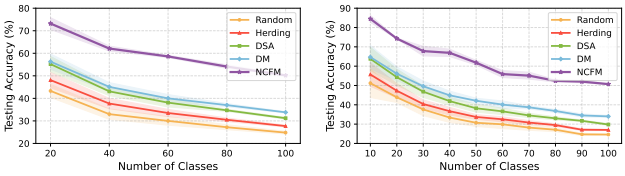
Dataset	CIFAR-10		CIFAR-100
IPC	500	1000	100
Random	73.2±0.3	78.4±0.2	42.8±0.3
MTT	↓ random	↓ random	49.2±0.4
TESLA	↓ random	↓ random	49.2±0.4
FTD	↓ random	↓ random	49.7±0.4
DATM	83.5±0.2	85.5±0.4	57.5±0.2
NCFM (Ours)	**84.0**±0.2	**86.8**±0.2	**58.8**±0.2
Whole Dataset	84.8±0.1		56.2±0.3

E. Continual Learning

In our study, we address the catastrophic forgetting problem in continual learning by employing NCFM. We store training samples in memory greedily while maintaining class balance. The model is retrained from scratch using only the latest memory, making the quality of memory construction crucial for continual learning performance. We conduct experiments on the CIFAR-100 dataset using both 5-step and

Table 9. Performance comparison (%) on ImageNet subsets with IPC=50.

Dataset IPC	ImageNette 50	ImageWoof 50	ImageFruit 50	ImageMeow 50	ImageSquawk 50	ImageYellow 50
RDED	83.8±0.2	61.5±0.3	-	-	-	-
NCFM (Ours)	**86.8**±0.5	**69.2**±0.8	**69.0**±0.7	**69.0**±0.4	**88.4**±0.4	**86.2**±0.6
Whole Dataset	87.4±1.0	67.0±1.3	63.9±2.0	66.7±1.1	87.5±0.3	84.4±0.6



(a) (b)

Figure 8. Performance evaluation of continual learning methods on CIFAR-100 with (a) 5-step and (b) 10-step settings.

10-step continual learning settings. In these settings, the dataset is divided into 5 and 10 stages, respectively, with each stage introducing new classes incrementally. Our experiments are repeated 5 times with varying class orders to ensure robustness. The memory budget is set to 20 images per class, consistent with previous studies. We compare our method against Random selection, DSA and DM. Our method synthesizes condensed images using the same hyperparameters as the CIFAR-100 experiment with 10 images per class, ensuring a fair comparison.

As illustrated in Figure 8a and Figure 8b, our method consistently outperforms the competitors in both settings, demonstrating the superior quality of the condensed data for continual learning. The final performance of our method reaches 50.2% in the 5-step setting and 50.7% in the 10-step setting, while DM achieves 33.8% and 34.0% respectively.

F. Further Results on Alternative Backbone Networks

In this section, we extend our analysis to include results on backbone networks beyond the previously used ConvNet. Specifically, we employed VGG-11, ResNet-18, and AlexNet as backbone architectures to perform dataset distillation. These networks were chosen to represent a diverse range of architectures, including shallow and deep models with varying capacities and inductive biases. This allows us to evaluate the generalizability and robustness of our distillation approach across different network designs.

The results, summarized in Table 10, demonstrate the effectiveness of our distillation method across all tested architectures.

Table 10. Cross-architecture generalization performance (%) on CIFAR-10, 50 IPC. The distilled data is trained on one architecture (T) and then evaluated on another architecture (E).

Method	T\E	ConvNet	AlexNet	VGG-11	ResNet-18
DC	ConvNet	53.9±0.5	28.8±0.7	38.8±1.1	20.9±1.0
CAFE	ConvNet	55.5±0.4	34.0±0.6	40.6±0.8	25.3±0.9
DSA	ConvNet	59.9±0.8	53.3±0.7	51.0±1.1	47.3±1.0
DM	ConvNet	65.2±0.4	61.3±0.6	59.9±0.8	57.0±0.9
KIP	ConvNet	56.9±0.4	53.2±1.6	53.2±0.5	47.6±0.8
MTT	ConvNet	71.6±0.2	48.2±1.0	55.4±0.8	61.9±0.7
FTD	ConvNet	73.8±0.2	53.8±0.9	58.4±1.6	65.7±0.3
DATM	ConvNet	76.1±0.3	45.0±0.7	59.4±0.6	66.3±0.1
ATT	ConvNet	74.5±0.4	60.0±0.9	61.7±0.9	66.3±1.1
IID	ConvNet	69.0±0.2	67.3±0.2	67.3±0.3	68.3±0.2
DataDAM	ConvNet	67.0±0.4	63.9±0.9	64.8±0.5	60.2±0.7
NCFM	ConvNet	**78.3**±0.3	**75.5**±0.3	**75.5**±0.3	**73.8**±0.2
	AlexNet	71.0±0.7	69.5±0.2	72.5±0.5	67.9±0.2
	VGG-11	69.4±0.6	69.6±0.3	71.7±0.4	69.8±0.2
	ResNet-18	69.4±0.4	68.9±0.3	71.1±0.5	71.2±0.5

G. Cross-architecture Performance on ViTs

To further validate the architectural generalization capabilities of our method, we conducted cross-architecture experiments on ImageNette and ImageWoof datasets using Vision Transformers (ViTs). We evaluated ViT on distilled images by ConvNets.As shown in Table 11, NCFM demonstrates superior performance over the previous SOTA method RDED, achieving remarkable accuracy gains of 17.2% and 11.9% on ImageWoof and ImageNette respectively at 10 IPC. The consistent performance elevation across different IPC settings underscores the robustness of NCFM during distillation.

Table 11. Cross-architecture generalization performance (%) on ImageNette and ImageWoof. The synthetic data is condensed using ConvNet, and evaluated on ViT.

Dataset	ImageNette		ImageWoof	
IPC	10	50	10	50
RDED	59.6±1.6	75.8±2.0	38.6±1.0	55.2±1.1
NCFM (ours)	**71.5**±1.1	**85.4**±1.0	**55.8**±0.9	**65.2**±0.9

H. Correlation between CFD and MMD

To better understand NCFM, we examine the relationship between the Characteristic Function Discrepancy (CFD) and Maximum Mean Discrepancy (MMD).

CF as Well-Behaved Kernels in the MMD Metric. The CF discrepancy term $\int_{\boldsymbol{t}} \sqrt{\text{Chf}(\boldsymbol{t}; f)} dF_{\mathcal{T}}(\boldsymbol{t})$ in our loss can be viewed as a well-behaved kernel in MMD, specifically as a *Characteristic Kernel*. Unlike MMD, which relies on fixed kernels, NCFM adaptively learns $F_{\mathcal{T}}(\boldsymbol{t})$, enabling flexible kernel selection for optimal distribution alignment. Furthermore, mixtures of Gaussian distributions within the CF framework produce well-defined characteristic kernels. When MMD employs a characteristic kernel of the form $\int_{\boldsymbol{t}} e^{-j\langle \boldsymbol{t}, \boldsymbol{x}-\tilde{\boldsymbol{x}} \rangle} dF_{\mathcal{T}}(\boldsymbol{t})$, it aligns with the structure of CFD, demonstrating that *MMD is a special case of CFD* when only specific moments are matched. This insight also explains the minimal memory overhead observed as IPC grows, highlighting the efficiency of our approach.

Computational Advantage of CFD over MMD. In contrast to MMD, which requires *quadratic* time in the number of samples for approximate computation, CFD operates in *linear* time relative to the sampling number of frequency arguments. This efficiency makes CFD substantially faster and more scalable than MMD, offering a particular advantage for large-scale datasets.

I. Phase & Amplitude Explanation

In this section, we conducted controlled experiments on MNIST through deliberate misalignment strategies. As illustrated in Figure 9, swapping amplitude spectra between class-specific distributions while preserving phase information (and vice versa) leads to significant semantic degradation in synthesized digits. Specifically, amplitude mismatch causes blurred digit contours, while phase misalignment disrupts structural coherence. These observations empirically validate our theoretical analysis in Section 4.2.1, where we demonstrate that amplitude primarily governs diversity through distribution scaling, while phase alignment encodes structural priors crucial for realism.

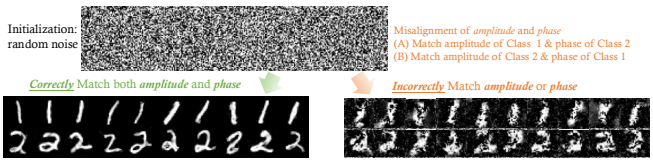


Figure 9. Phase and amplitude alignment analysis on MNIST. We conducted experiments on MNIST by combining *amplitude* from one set (*e.g.*, class 1 or class 2) with *phase* from another (*e.g.*, class 2 or class 1).

J. Visualization Comparison between MMD-based method and NCFM.

We provide additional visualization comparisons between NCFM and MMD-based DM on the challenging ImageFruit dataset in Figure 10. Our method generates visually distinct specimens with crisper fruit boundaries and better color gradient preservation. The improved perceptual quality directly correlates with NCFM's explicit optimization of both magnitude and angular components in the characteristic function space, enabling a more precise recovery of high-frequency details.

Figure 10. Visualization comparison between DM and NCFM on ImageFruit dataset with 10 IPC.