Robust Dataset Condensation using Supervised Contrastive Learning

Nicole Hee-Yeon Kim and **Hwanjun Song**

Korea Advanced Institute of Science and Technology (KAIST)

Department of Industrial Systems Engineering

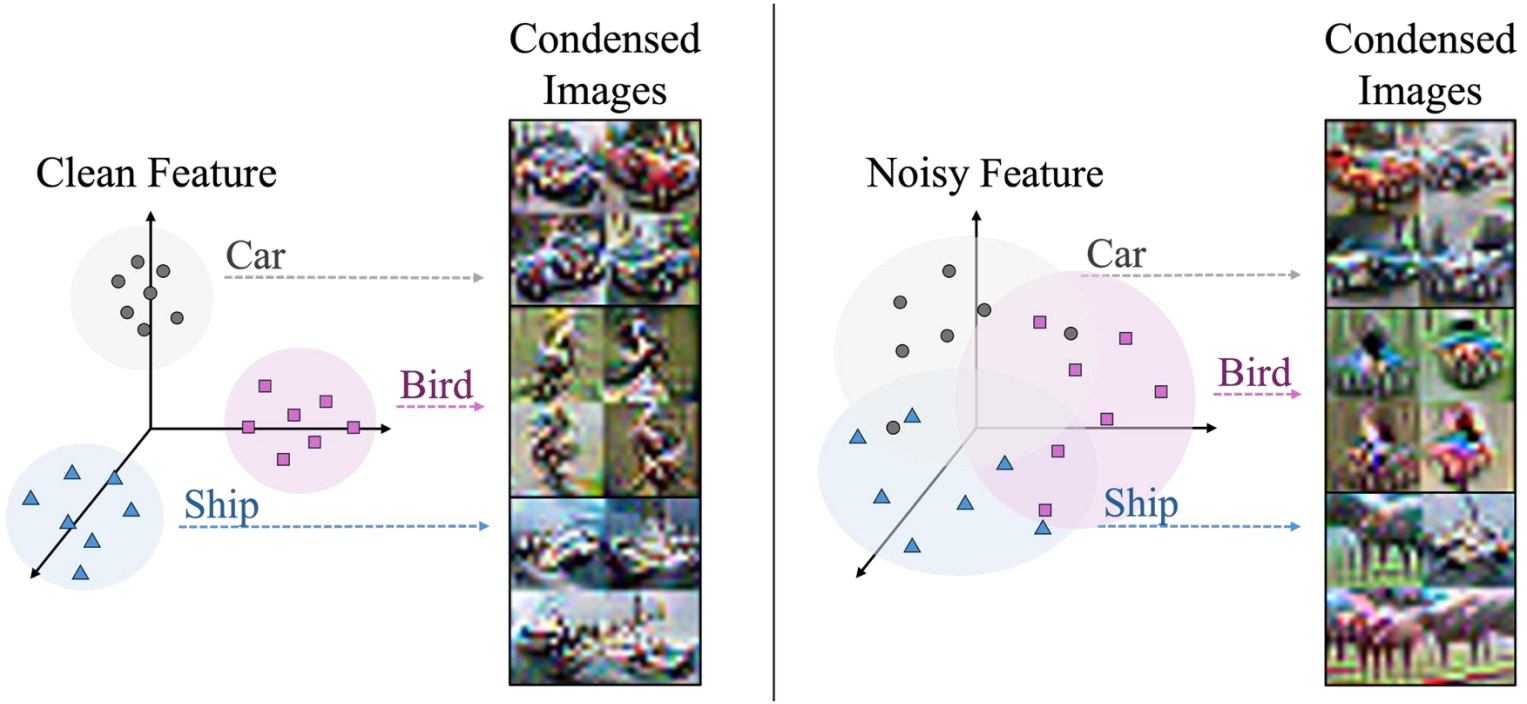
Data Intelligence System Lab (DISL.)

ICCV 2025, Honolulu, Hawaii

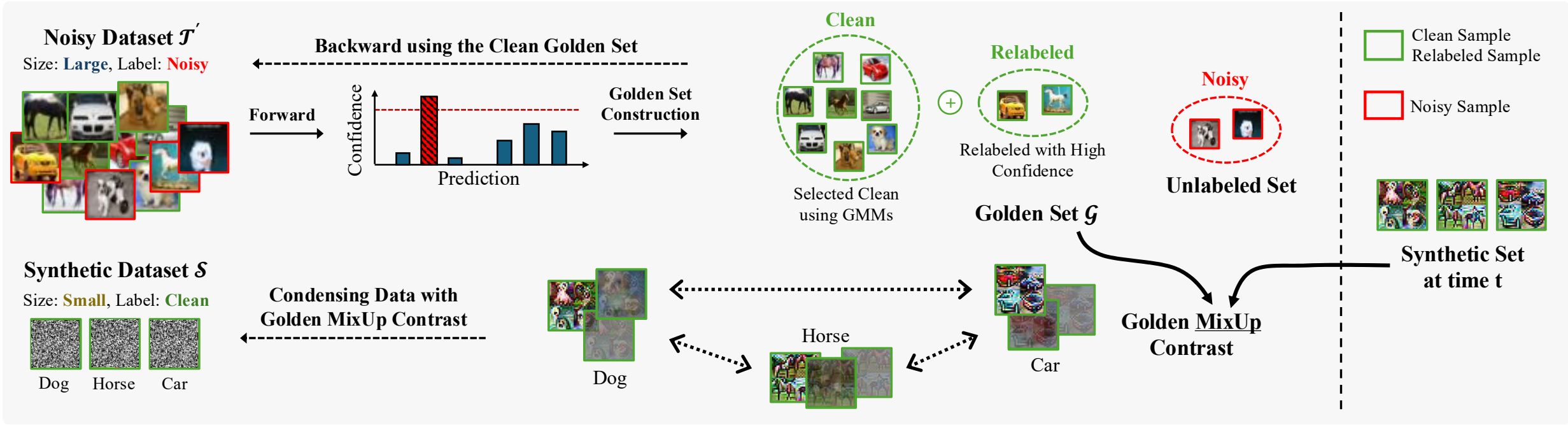
Dataset Condensation: Concept and Limitations

- Dataset condensation compresses large-scale datasets into a compact synthetic set that preserves essential learning signals.

- However, existing methods fail under noisy labels, producing distorted and unreliable synthetic data.
- For example, under a 40% asymmetric noise environment, the ship class is distorted into a form with horse legs (image on the right).

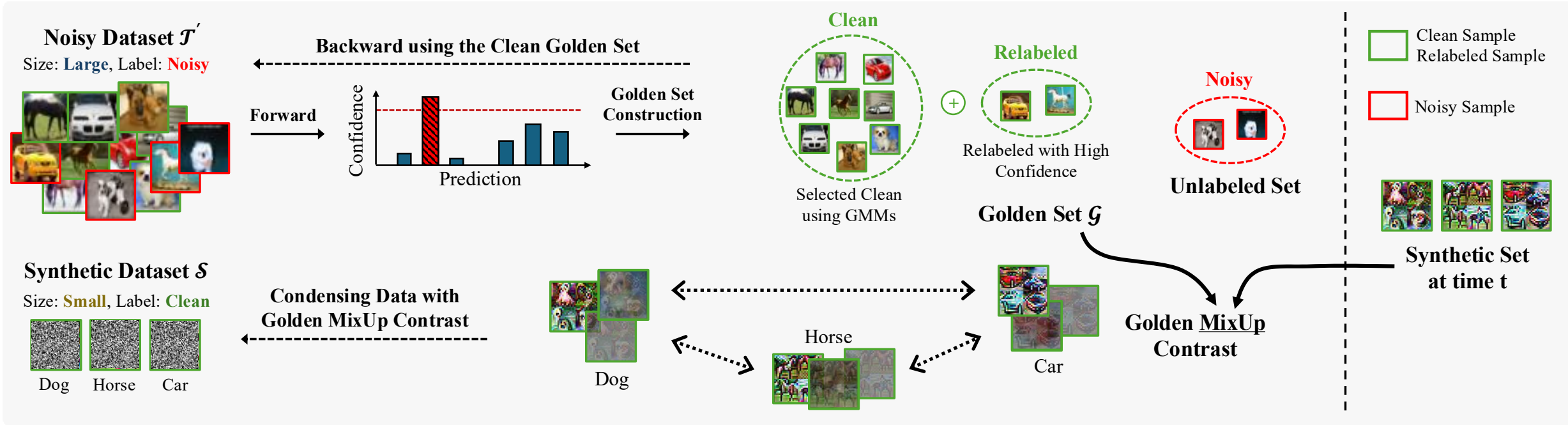


Our Solution: Robust Dataset Condensation (RDC)



- We propose Robust Dataset Condensation (RDC), the first end-to-end framework resilient to noisy labels.
- RDC leverages supervised contrastive learning and Golden MixUp Contrast to generate clean, robust synthetic datasets.

How RDC Works



- A golden set of clean and relabeled samples is extracted from noisy data and combined with synthetic samples.
- Golden MixUp Contrast transfers reliable signals from the golden set, enhancing diversity and suppressing noise.

Performance under Noisy Labels

CIFAR-10	Clean		Asymmetric Noise				Symmetric Noise				Real-world Noise			
Noise Ratio	≈ 0%		20%		40%		20%		40%		Random1 (17%)		Worse (40%)	
Img/Cls	10	50	10	50	10	50	10	50	10	50	10	50	10	50
Random	22.28	36.05	19.55	29.12	19.53	24.48	17.11	29.31	18.36	24.28	20.19	31.88	20.19	28.26
IDM	45.10	60.24	39.63	47.64	30.61	34.85	42.97	56.54	41.17	45.62	44.68	57.19	38.36	49.46
IDM + Two-stage	43.77	60.24	45.59	60.23	33.29	38.61	45.70	59.27	**46.52**	59.26	45.25	60.15	45.29	60.10
IDM + RDC (Ours)	**47.28**	**60.80**	**46.92**	**61.76**	**41.35**	**55.55**	**47.12**	**63.15**	46.23	**59.61**	**48.36**	**61.85**	**46.52**	**61.93**
Whole Dataset	95.37		81.42		58.81		84.00		64.79		85.45		67.36	

- RDC consistently outperforms existing methods across noise types and ratios.
- In CIFAR-10 with 40% asymmetric noise, RDC achieves 55.55%, far higher than the two-stage method (38.61%) and close to the clean upper bound (60.24%).
- The improvement generalizes to CIFAR-100 and Tiny-ImageNet, confirming RDC's robustness and scalability.

Why It Works: Component Analysis

Component	Asymmetric	Symmetric	Real-world
IDM	30.61	41.17	38.36
+ (1) SSL (DivideMix)	35.23	41.93	44.47
+ (2) SupCon wo. Augment	33.89	43.63	44.99
+ (3) SupCon w. Augment	35.24	44.94	45.58
+ (4) Golden MixUp Contrast	41.35	46.23	46.52

- Adding semi-supervised learning (DivideMix) and SupCon yields moderate gains, but limited under asymmetric noise.
- Simple augmentation (flip, crop) fails to resolve the diversity problem of synthetic sets.
- Golden MixUp Contrast (GMC) provides the largest boost, proving essential for robust and diverse condensation.

Seeing the Difference: Visualization of Condensed Datasets

- Baseline condensation (Acc-DD) suffers severe class interference, mixing features across categories.
- Two-stage cleaning reduces interference but still leaves residual contamination.
- RDC fully removes cross-class noise, yielding clean and accurate synthetic representations of horse and ship.

Key Takeaways

- RDC is the first dataset condensation method robust to noisy labels.
- Supervised contrastive learning separates classes, reducing cross-class interference.
- Golden MixUp Contrast transfers reliable signals from real data, enhancing diversity and stability.
- RDC shows strong robustness and generalizability across noise types, levels, and architectures.


ICCV
OCT 19-23, 2025
HONOLULU
HAWAII