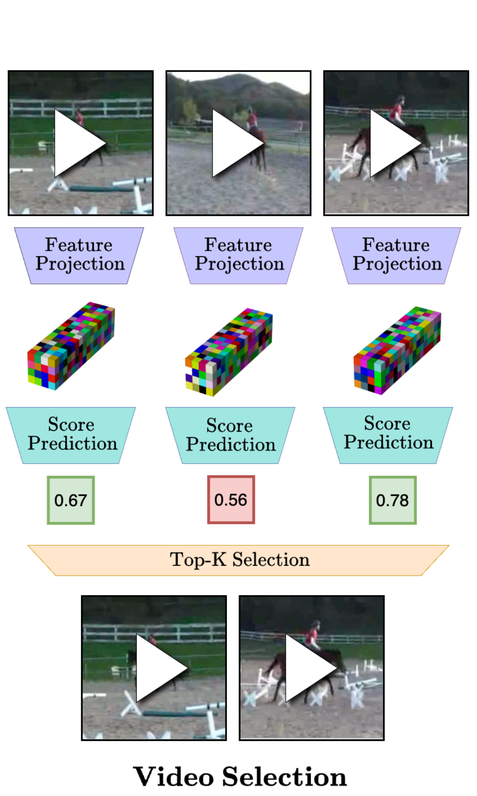
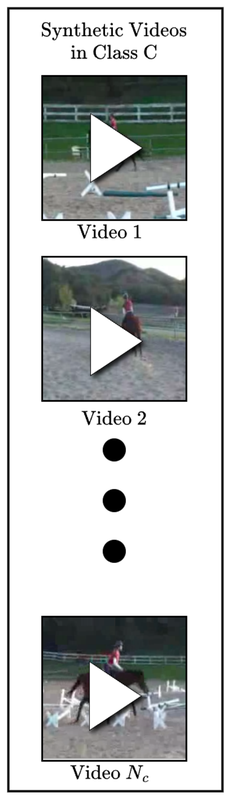
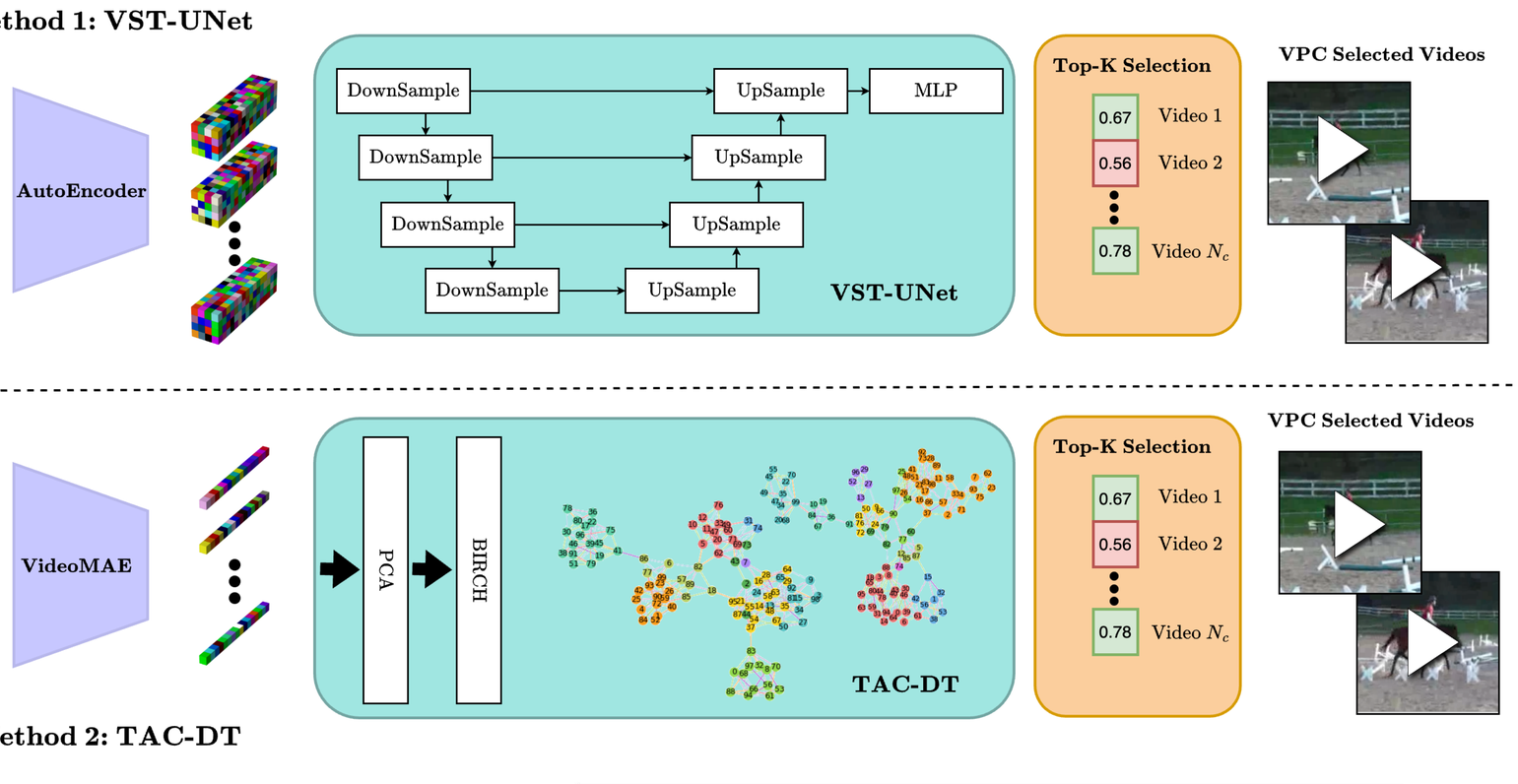
Video Dataset Condensation with Diffusion Models

Zhe Li, Hadrien Reynaud, Mischa Dombrowski, Sarah Cechnicka,
Franciskus Xaverius Erick, Bernhard Kainz







Introduction & Challenges

Dataset distillation: The goal

- Condense a large-scale real dataset into a significantly smaller synthetic dataset.
- Achieve comparable performance to the original large dataset.

Challenges:

- Significantly reduce the data volume and computation cost.
- Handling temporal information in videos.

Diffusion Model Utilization:

- Train a class-conditional diffusion model.
- Generate diverse synthetic video data only once and store them.

Algorithm 1 Our proposed algorithm for video distillation

Input: synthetic videos $\mathcal{S}_g$, VPC;
Initialize the selection VST-UNet model M_{VST}
for E_s epochs **do**
  for class c **do**
    $V_c \leftarrow \mathcal{S}_g$
    $\hat{V}_c = \text{AutoEncoder}(V_c)$
    $f_c,\ p_c = M_{VST}(\hat{V}_c)$
    $L = L_{ce} + \lambda_d L_{div} + \lambda_r L_{rep}$
  end for
end for
for class c **do**
  $f_c,\ p_c = M_{VST}(\hat{V}_c)$
  $S_c = V_c\,[argsort(p_c)[:\text{VPC}]]$
end for
Train classifiers on the selected synthetic videos S.

$$L_{div} = \frac{1}{|\mathcal{S}_c|(|\mathcal{S}_c|-1)} \sum_{t\in\mathcal{S}_c} \sum_{\substack{i\in\mathcal{S}_c \\ i\neq t}} d(x_t, x_i)$$

$$L_{rep} = \exp\left(-\frac{1}{N_c}\sum_{t=1}^{N_c} \min_{i\in\mathcal{S}_c} \| x_t - x_i \|_2\right)$$

Ablation Study

- **Dataset Scope**: Whole UCF101 dataset, includes low-performing classes.
- **Clustering methods**: Random, MAE-Score, CSTA, DM-MMD.
- **Classifiers**: C3D and SwinTransformer.
- **Loss functions** in VST-UNet: CE, Div, and Rep losses.
- **Computational Efficiency**: The computation times are reduced.

		GPU	VPC	UCF101	MiniUCF	HMDB51	Kinetics400	SSv2
Diffusion	Time	4	–	–	–	~ 3 days	~ 7 days	~ 6 days
	Steps	–	–	–	–	500k	1M	1M
Distill	VST-UNet	1	1	3.5 h	1.5 h	2.5 h	20 h	16 h
	TAC-DT	1	1	2.5 h	1 h	1.5 h	17 h	14.8 h
	VST-UNet	1	5	11.5 h	6 h	6 h	67 h	33.5 h
	TAC-DT	1	5	7 h	2.5 h	2.7 h	38 h	25.5 h

Conclusion

1. Two video selection strategies:
   - VST-UNet and TAC-DT.
2. Significantly improvement:
   - Test accuracy and computational efficiency.

Datasets	MiniUCF		HMDB51		Kinetics400(top5)		SSv2(top5)	
VPC	1	5	1	5	1	5	1	5
Full real	60.14±0.95		36.24±0.83		34.6±0.5		29.0±0.6	
Full syn (100)	56.22±0.44		31.05±0.30		-		14.56±0.15	
DM [25, 30]	15.3±1.1	25.7±0.2	6.1±0.2	8.0±0.2	6.3±0.0	9.1±0.9	3.6±0.0	4.1±0.0
MTT [1, 25]	19.0±0.1	28.4±0.7	6.6±0.5	8.4±0.6	3.8±0.2	9.1±0.3	3.9±0.1	6.3±0.3
FRePo [25, 34]	20.3±0.5	30.2±1.7	7.2±0.8	9.6±0.7	-	-	-	-
DM+VD [25]	17.5±0.1	27.2±0.4	6.0±0.4	8.2±0.1	6.3±0.2	7.0±0.1	4.0±0.1	3.8±0.1
MTT+VD [25]	23.3±0.6	28.3±0.0	6.5±0.1	8.9±0.6	6.3±0.1	11.5±0.5	5.5±0.1	**8.3**±0.2
FRePo+VD [25]	22.0±1.0	31.2±0.7	8.6±0.5	10.3±0.6	-	-	-	-
VST-UNet (ours)	**26.28**±1.07	**41.81**±0.99	13.59±0.33	18.05±0.75	**6.67**±0.11	**11.82**±0.10	5.25±0.16	8.22±0.14
TAC-DT (ours)	25.20±0.41	41.47±0.32	**13.82**±0.33	**19.62**±0.28	6.50±0.19	11.72±0.35	**5.54**±0.19	8.06±0.17

Acknowledgments:
High-performance computing resources were provided by the Erlangen National High Performance Computing Center (NHR@FAU) at Friedrich-Alexander-Universit.t Erlangen-Nürnberg (FAU), under the NHR projects b143dc and b180dc. NHR is funded by federal and Bavarian state authorities, and NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. Additional support was received by the ERC - project MIA-NORMAL 101083647, DFG 513220538, 512819079, and by the state of Bavaria (HTA). S. Cechnicka is supported by the UKRI Centre for Doctoral Training AI4Health (EP / S023283/1).