UniSTD: Towards Unified Spatio-Temporal Learning across Diverse Disciplines

Supplementary Material

1. Implementation Details

Architecture. More details about the architectures of UniSTD are shown in Tab. 3. The decoder blocks are the same as the Encoder blocks but with the `TransposedConv2d` layers to replace the `Conv2d` layers in the Encoder layers for upsampling the features.

Training Details. We train the model for 90 epochs with the AdamW optimizer for the weights and SGD optimizer for the trainable rank of MoE. For the Adam optimizer, the weight decay is set to $1e-5$, the learning rate is set to $7e-4$ with cosine learning rate scheduler and the first 5 epochs are used for warm-up (using $1e-8$ learning rate). For the SGD optimizer, we enable the Nesterov momentum and set the learning rate to 0.05.

2. Additional Results

Task	Model	PSNR (↑)	SSIM (↑)
BAIR	Ours	20.3	0.86
	TAU	19.8	0.86
	SimVPV1	20.3	0.86
	SimVPv2	19.9	0.85
KTH	Ours	28.4	0.92
	TAU	27.9	0.90
	SimVPv1	27.7	0.90
	SimVPv2	27.9	0.90
MMNIST	Ours	20.5	0.90
	TAU	18.9	0.85
	SimVPv1	19.5	0.88
	SimVPv2	19.0	0.85

Table 1. Task-wise comparison of our unified model and the single task baselines.

More Evaluation Metrics. We provide an additional evaluation metric RMSE in Tab. 5, one can see that our method still yield best performance across various methods. .

Task-wise Training Results. In Tab. 1, we show the additional results of our joint trained model compared to the single-task training (independent training) of baselines. Our model shows significant improvements on metrics of both PSNR and SSIM, this further indicates that the joint training can benefit the learning process of each task. We train the baseline using our training settings for fair comparison.

Efficiency Analysis. In Tab. 4 and Tab. 2, we show the computational complexity (FLOPs) and number of trainable parameters of the proposed method and baseline, respectively. On the one hand, long-range spatial modeling of Transformer allows UniSTD to use much smaller spatial dimensions, thus more efficient on mid/large resolution tasks (e.g., SEVIR, Human, etc) in terms of FLOPs. On the other hand, our method uses only about 50% trainable parameters of the baselines while achieving much better performance.

Task	Method	Trainable Params. (↓)
KITTI+Traffic.+KTH	SimVPv1	45.8M
	SimVPv2	35.3M
	TAU	33.9M
	UniSTD (Ours)	18.5M
BAIR+TaxiBJ+Human +City.+KTH+Traffic. +KITTI	SimVPv1	61.7M
	SimVPv2	47.8M
	TAU	45.8M
	UniSTD (Ours)	23.6M

Table 2. Number of trainable Parameters of UniT and baseline.

References

[1] Yuan Yuan, Jingtao Ding, Jie Feng, Depeng Jin, and Yong Li. Unist: A prompt-empowered universal model for urban spatio-temporal prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4095–4106, 2024. 2

Table 3. Architecture details of UniSTD.

	TaxiBJ	Traffic4Cast	MMNIST	BAIR	Human3.1M	KTH	Cityscapes	KITTI	SEVIR	ENSO
GFLOPs	6.91	26.43	7.36	139.97	31.53	29.89	85.01	37.78	88.76	62.97
Shape: (B*T, C, H, W)	(B*4, 2, 32, 32)	(B*9, 8, 128,112)	(B*10, 1, 64, 64)	(B*2, 3, 64, 64)	(B*4, 3, 256, 256)	(B*10, 1,128, 128)	(B*2, 3, 128, 128)	(B*10, 3, 128, 160)	(B*13, 1, 384, 384)	(B*12, 1, 24, 48)
Encoder Block: – `Conv2d` (stride=1, channels=T') – `GroupNorm` – `SiLU` – `Conv2d` (stride=2, channels=C') – `GroupNorm` – `SiLU`	*2	*3	*3	*2	*4	*3	*3	*3	*4	*1
					C'=64					
Shape: (B, T*C', H', W')	(B, 256, 8, 8)	(B, 576, 16, 14)	(B, 640, 8, 8)	(B, 128, 16, 16)	(B, 256, 16, 16)	(B, 640, 16, 16)	(B, 128, 16, 16)	(B, 640, 16, 20)	(B, 832, 24, 24)	(B, 768, 12, 24)
Projection Layer: – `Conv2d` (stride=1, channels=768)	*1	*1	*1	*1	*1	*1	*1	*1	*1	*1
Shape: (B, 768, H', W')	(B, 768, H', W')									
– `Reshape`	N=H'*W', L=768									
Shape: (B, N, L)	(B, 64, 768)	(B, 224, 768)	(B, 64, 768)	(B, 256, 768)	(B, 256, 768)	(B, 256, 768)	(B, 256, 768)	(B, 320, 768)	(B, 576, 768)	(B, 288, 768)
Backbone: – `Transformer Blocks (shared)` with Rank-Adaptive MoE and Temp. Attn.	* 12									

Table 4. GFLOPs (lower is better) comparison.

	TaxiBJ	Traffic4Cast	MMNIST	BAIR	Human3.1M	KTH	Cityscapes	KITTI	SEVIR	ENSO
UniSTD (Ours)	6.91	**26.43**	**7.36**	**139.97**	**31.53**	**29.89**	85.01	**37.78**	**88.76**	62.97
SimVP_{v1}	3.53	40.01	15.15	277.64	231.80	50.06	127.70	62.60	426.92	25.58
SimVP_{v2}	2.54	32.06	12.28	194.77	166.59	64.26	88.79	80.35	398.34	22.82
TAU	**2.43**	30.62	11.75	185.44	158.83	61.25	**84.78**	76.59	379.91	21.79

Table 5. RMSE (↓) metrics.

	TaxiBJ	Traffic4Cast	MMNIST	BAIR	Human3.1M	KTH	Cityscapes	KITTI
UniSTD **(Ours)**	0.54	8.78	6.27	11.54	10.46	5.42	10.82	39.49
$\text{SimVP}_{\text{3 task}}$	2.86	-	-	18.79	35.45	-	-	-
$\text{TAU}_{\text{3 task}}$	1.69	-	-	-	34.48	-	26.56	76.59
UniST [1]	0.90	-	-	25.76	-	23.36	-	56.95