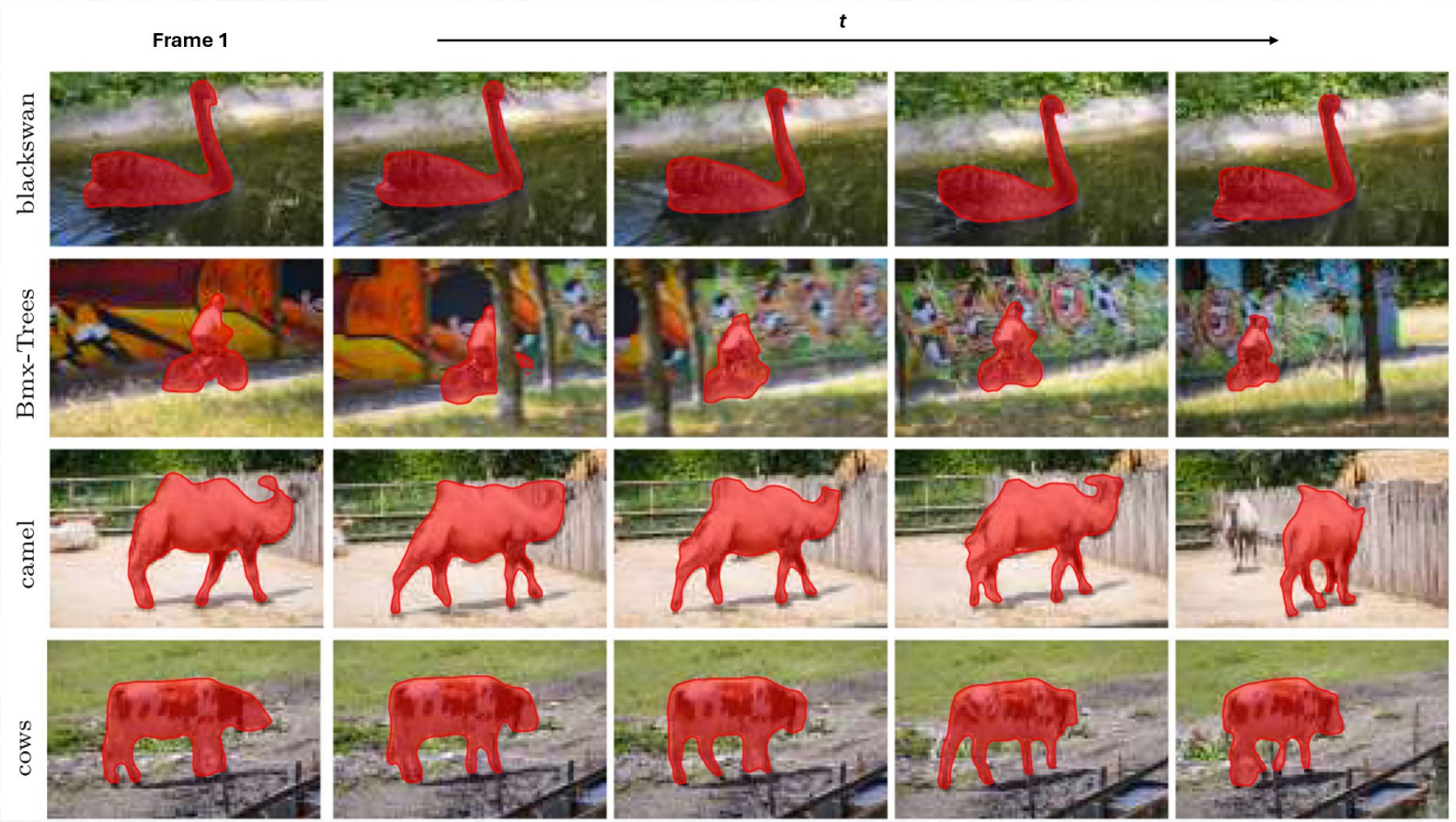
Betrayed by Attention: A Simple yet Effective Approach for Self-supervised Video Object Segmentation *[Submitted on 29 Nov 2023]*

Shuangrui Ding, Rui Qian, Haohang Xu, Dahua Lin, Hongkai Xiong

Eva Mičánková Lenka Šoková David Kedra

28.3.2024

Segmentation tracking in video



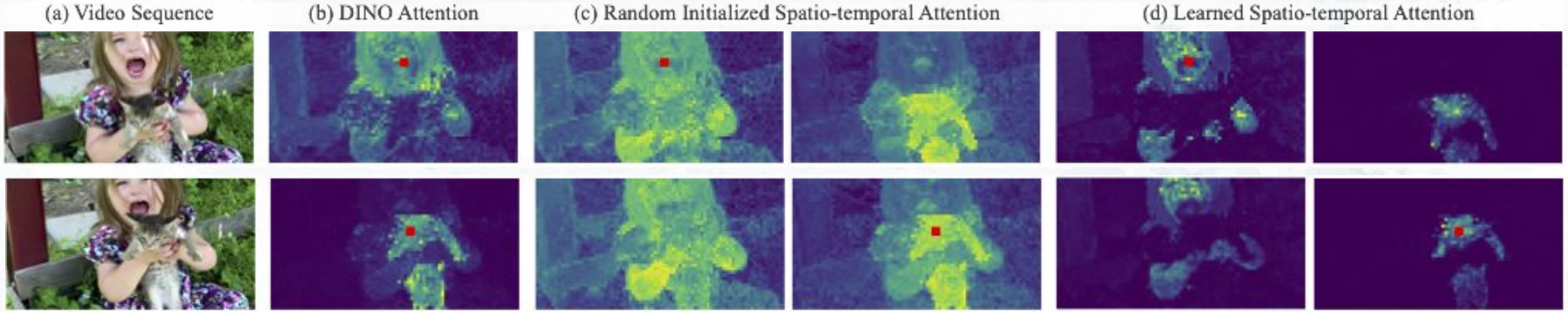
O modelu

- Vychází ze self-supervised předtrénovaných vision transformerů (DINO ViT) pro extrakci pokročilých obrazových příznaků
- Analýza attention patternů transformerů ukázala, že z nich lze extrahovat i informace o jednotlivých objektech
- Lze tyto atributy využít pro segmentaci a sledování objektů ve videu?

- Model informace dále zpracovává “spatio-temporal transformer bloky”, které trénuje tak, aby vyčistil oblasti od šumu a následně shlukoval do segmentů objektů

- Dle slov autorů je zajímavé, že tato naivní metoda shlukování přináší neočekávaně konkurenceschopné výsledky

“Attention leaks the object’s position”



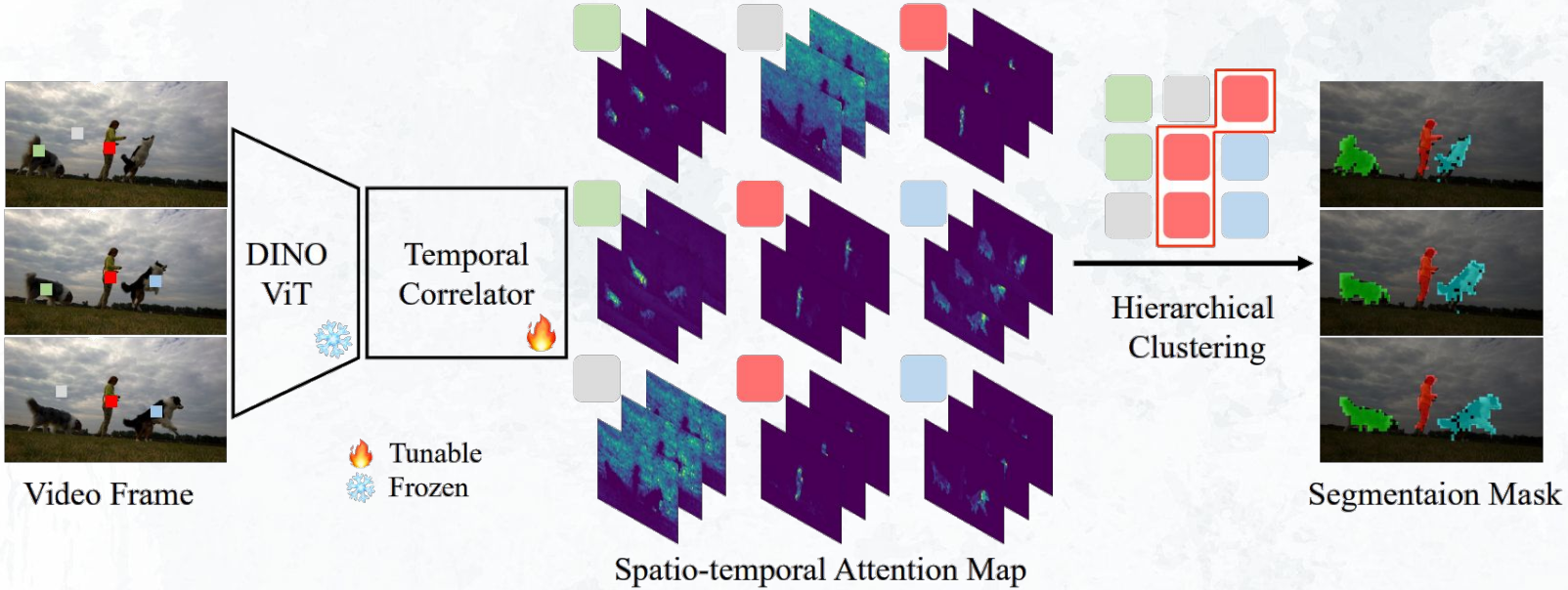
(a) Vstupem jsou pouze RGB snímky

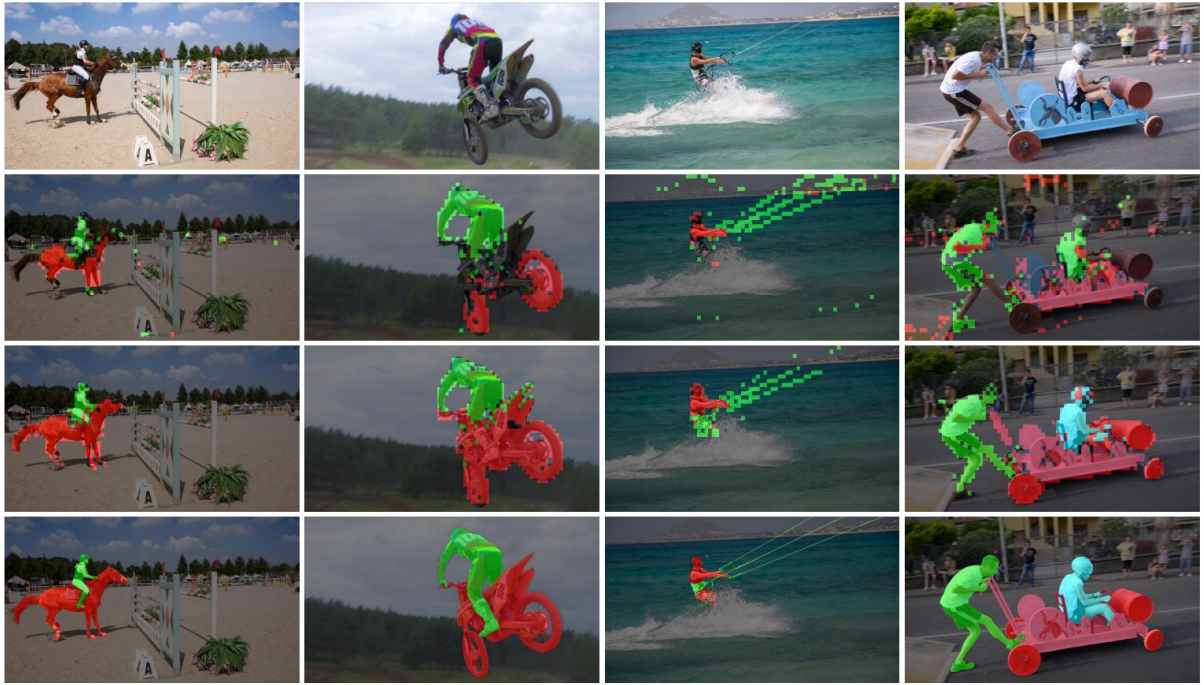
(b) Attention mapy patternů z předtrénovaných transformerů DINO

(c) Náhodně inicializovaný “spatio-temporal transformer block”

(d) Mapy natrénovaných bloků s redukovaným šumem

Architektura modelu





	MOVi-E		YTVIS-19		DAVIS-17			
Model	FG-ARI	mIoU	FG-ARI	mIoU	FG-ARI	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
SAVi [27]	42.8	16.0	11.1	12.7	-	-	-	-
STEVE [57]	50.6	26.6	20.0	20.9	-	-	-	-
OCLR [64]	-	-	15.9	32.5	14.7	-	34.6	-
VideoSAUR [73]	73.9	35.6	39.5	29.1	-	-	-	-
SOLV [1]	80.8	-	29.1	45.3	32.2	-	30.2	-
SMTC [53]	-	-	31.4	38.8	33.3	40.5	36.4	44.6
TimeT* [54]	-	-	37.9	40.4	35.5	40.0	35.8	44.2
Ours	83.4	40.2	44.3	50.1	40.1	43.9	39.2	48.6
Ours†	84.4	40.7	44.5	50.1	41.6	43.7	39.4	48.0


RGB
Ours
RGB
Ours