Implicit texture mapping for multi-view video synthesis

Mohamed Ilyes Lakhal*, Oswald Lanz, Andrea Cavallaro
mohamed.ilyes.lakhal@huawei.com

1. Introduction

Goal

- Generate a video of human motion from a different view-point

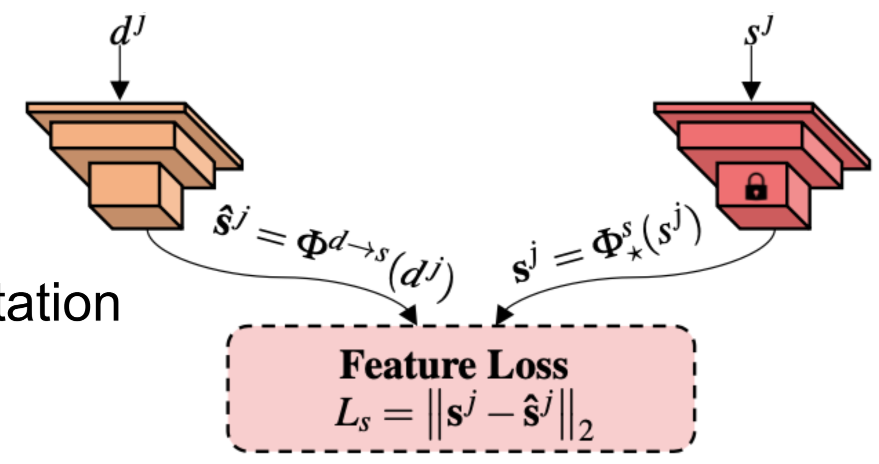
Approach

- Motion learning

through texture-less representation

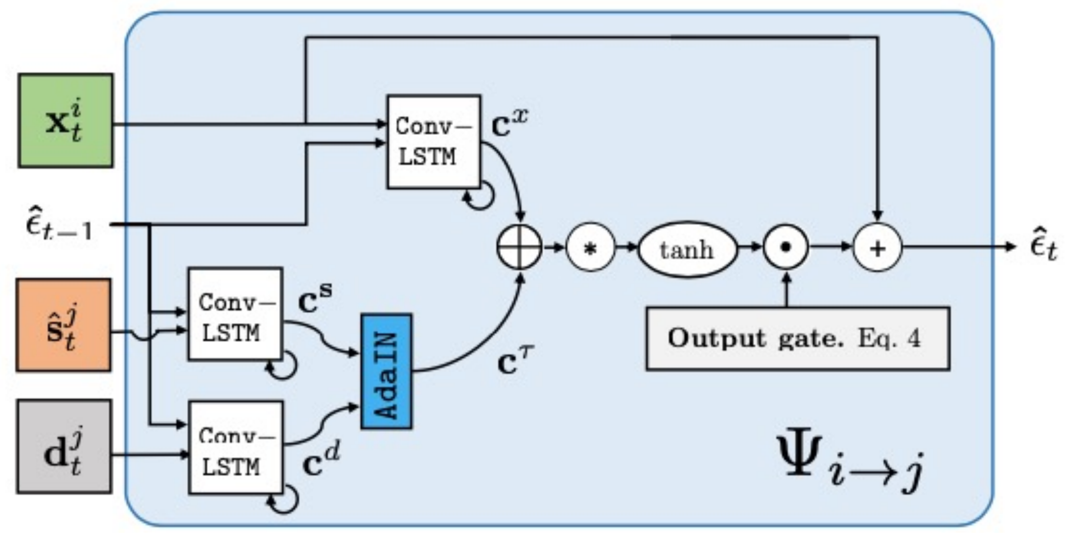
Applications

- data augmentation



2. VA-LSTM

- Target-view learning through Conv-LSTM
- Incorporates the texture-less representation



3. View Adaptive Network (VA-Net)

- VA-LSTM for target-view feature approximation
- Two-stage pipeline for synthesis and refinement

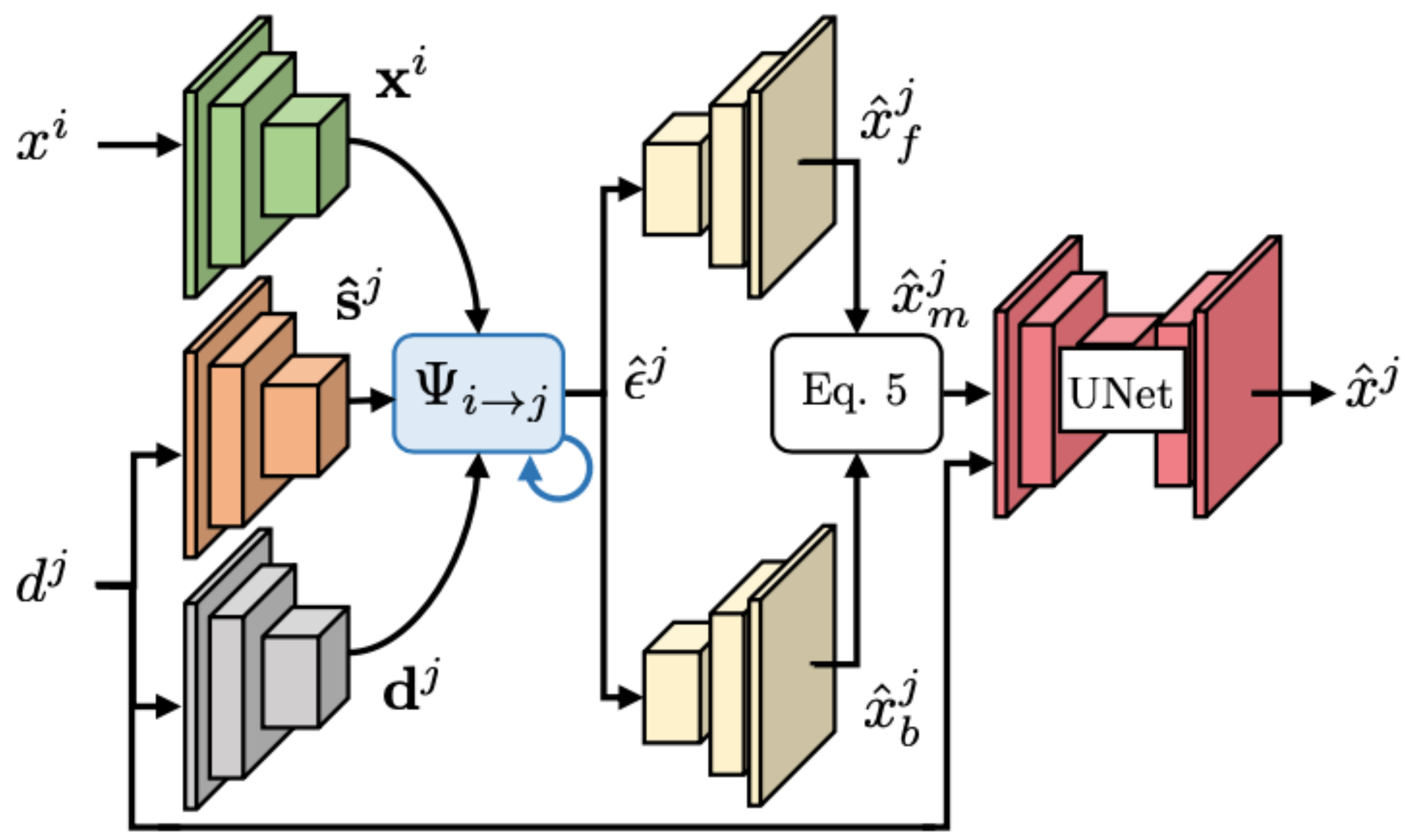
Motion network:

- Separate foreground and background synthesis
- Foreground feature estimation using the teacher-student approach (see Sec. 3)

Refinement network:

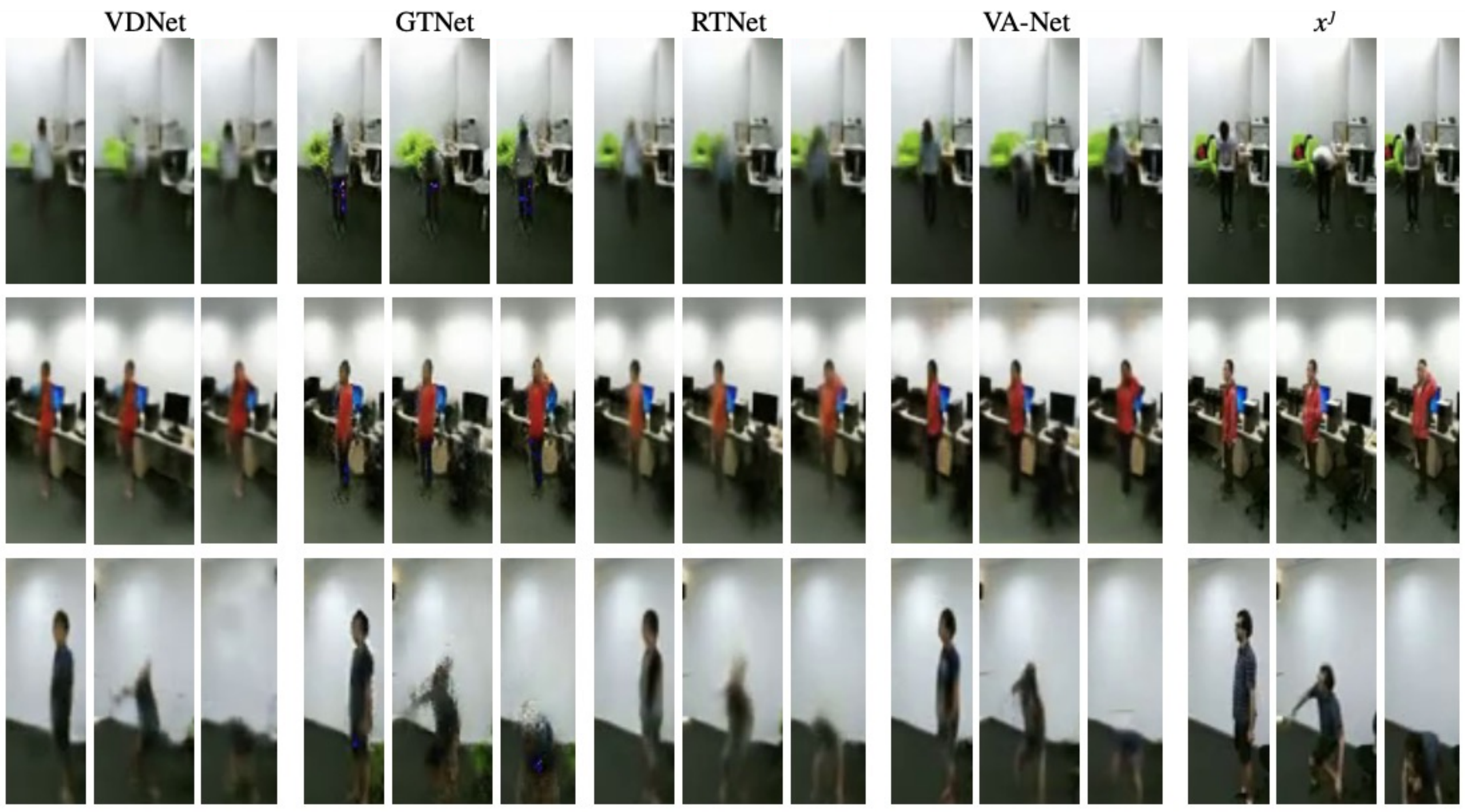
- UNet-like network to retain spatial information
- Use explicit edge-loss (Sobel-filter) penalization to encourage high-frequency details:

$$L_e = \|\mathbf{C}_x * x^j - \mathbf{C}_x * \hat{x}^j\| + \|\mathbf{C}_y * x^j - \mathbf{C}_y * \hat{x}^j\|$$



4. Results

Comparison with state-of-the-art methods



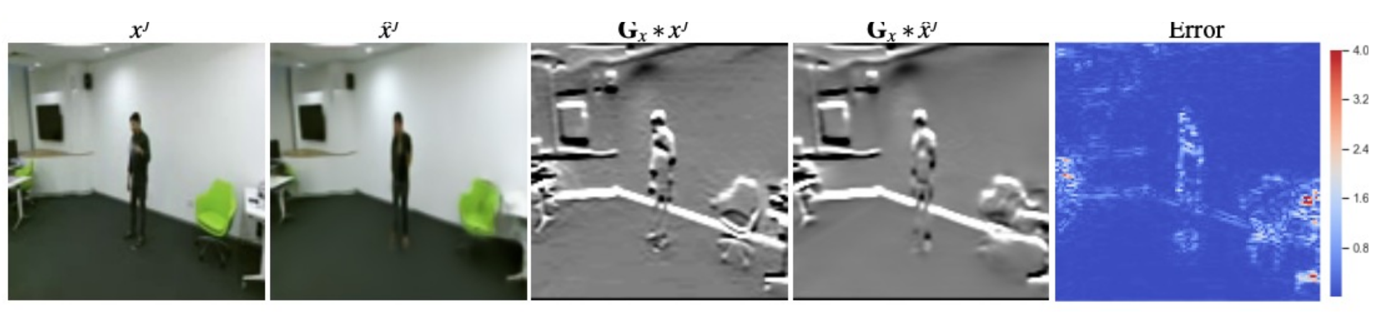
Method	Cat.	$\Psi_{i\to j}$	Modality	SSIM	M-SSIM	PSNR	M-PSNR	FVD	L_2	PCK 0.20	PCK 0.05	PCK 0.01	Precision	Recall	F1
PG2	Image	CNN	s^j	.582	.954	16.90	25.87	11.84	10.91	97.8	74.7	14.4	88.1	12.8	22.4
PATN	Image	CNN	s^i, s^j	.534	.948	16.24	24.55	13.11	11.68	98.0	69.7	10.2	88.4	.1	17.8
XingGAN	Image	CNN	s^i, s^j	.445	.933	13.32	23.29	14.47	26.41	89.6	17.8	.01	84.5	.1	.1
VDNet	Gen	RNN	d^j, s^j	.821	.972	23.18	29.70	5.78	4.37	99.3	92.4	51.2	91.0	55.3	68.7
Baseline	Gen	CNN	d^j	.813	.965	22.82	27.87	5.28	5.80	99.4	89.4	41.5	**92.5**	30.6	46.0
RTNet	Motion	RNN*	$x^j_{t=1}$	.933	–	29.07	–	–	–	–	–	–	–	–	–
	Motion	RNN	$\hat{x}^j_{t=1}$	.878	–	25.27	–	–	–	–	–	–	–	–	–
	Motion	RNN	$d^j_{t=1}$	.887	.977	25.76	30.68	4.14	4.13	99.4	93.0	53.4	91.7	56.6	70.0
GTNet	Two-stream	CNN	$\mathcal{T}^j, \mathcal{S}^j, d^j$	.823	.981	23.81	32.50	4.96	3.95	99.5	93.0	57.6	92.3	52.7	67.1
VA-Net (motion)	Two-stream	RNN	$\mathcal{S}^j, d^j$	.845	.980	24.50	31.70	**3.63**	**2.87**	99.5	**95.5**	**67.7**	91.1	**69.9**	**79.1**
	Two-stream	RNN	d^j	.862	.978	24.71	31.15	3.70	3.75	99.4	93.1	58.6	91.9	58.3	71.3
VA-Net (refinement)	Two-stream	–	–	**.895**	.979	25.48	31.36	3.67	3.46	**99.6**	94.3	59.4	91.2	65.3	76.1

View-LSTM vs. VA-LSTM

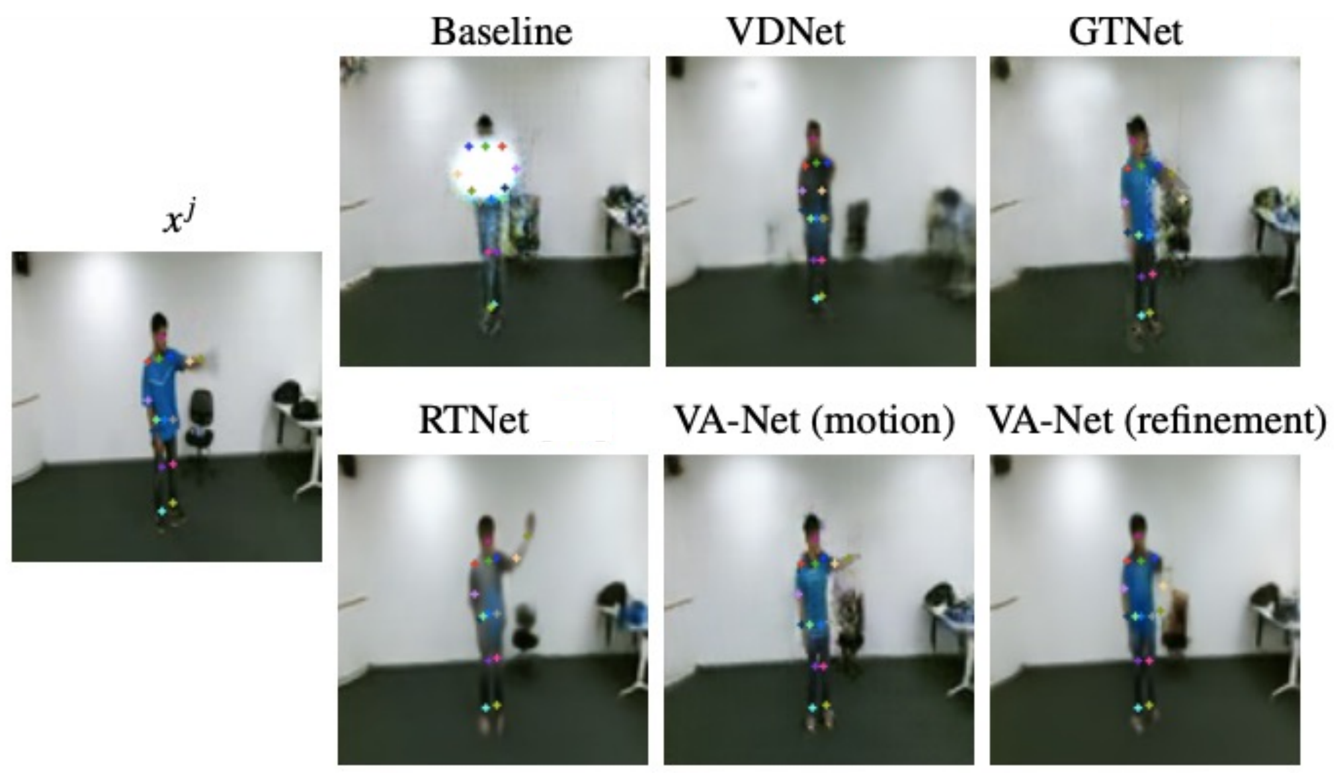
Model	Modalities	SSIM	M-SSIM	PSNR	M-PSNR
View-LSTM	d^j, s^j	.821	.972	23.18	29.70
	$d^j, \mathcal{S}^j$	.833	.975	23.44	30.35
VA-LSTM	d^j, s^j	.830	.976	23.78	30.89
	$d^j, \mathcal{S}^j$	.845	.980	24.50	31.70

	Method	SSIM	M-SSIM	PSNR	M-PSNR	#param
	View-LSTM	.821	.972	23.18	29.70	(14.36 + 22.41) M
VA-LSTM	+ AdaIN	.862	.978	24.71	31.15	13.70 M
VA-LSTM	+ branch	.851	.979	24.14	31.63	15.47 M
VA-LSTM	+ conv	.847	.979	24.42	31.65	13.83 M
VA-LSTM	+ sum	.842	.979	24.37	31.63	13.70 M

Edge-loss



Pose estimation results



References

[1] Lakhal M, Lanz O, Cavallaro A. View-LSTM: novel-view video synthesis through view decomposition. ICCV 2019
[2] Schatz K, Quintanilla E, Vyas S, and S Rawat Y. A Recurrent Transformer Network for Novel View Action Synthesis. ECCV 2020