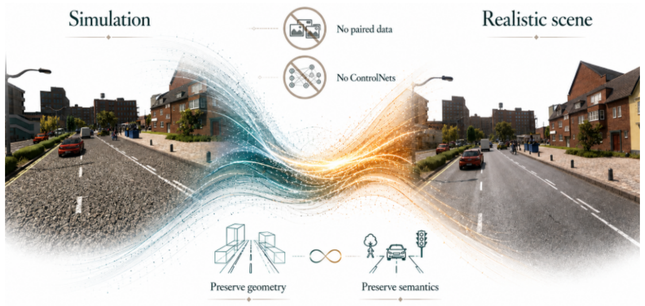
Wavelet Phase Diffusion for Structurally and Semantically Consistent Sim-to-Real Translation

Kaiwen Wang[1] Frank Bieder[2] Yinzhe Shen[1] Carlos Fernandez[1] Jan-Hendrik Pauls[1] Ömer Şahin Taş[2]
[1]Karlsruhe Institute of Technology (KIT) [2]FZI Research Center for Information Technology

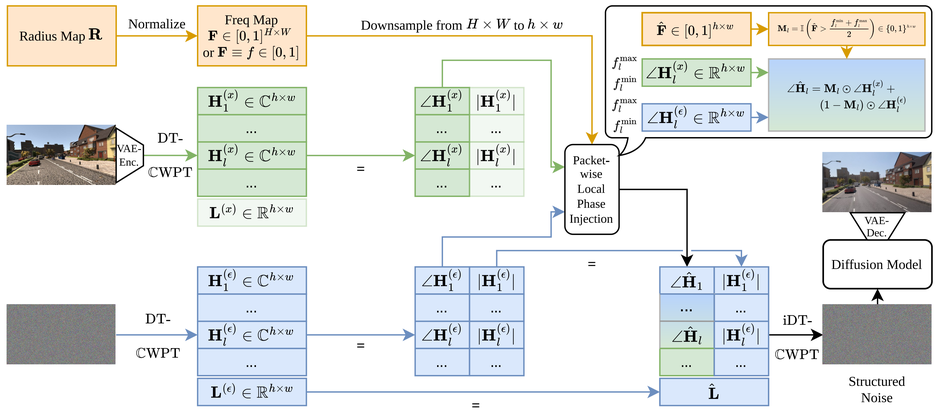
WACV JAN 4-8 2027 Orlando, Florida

Concept

How can we translate simulated scenes into realistic ones without **paired data** or **control modules**, while preserving **geometric and semantic** consistency?



Method



Why Wavelet?

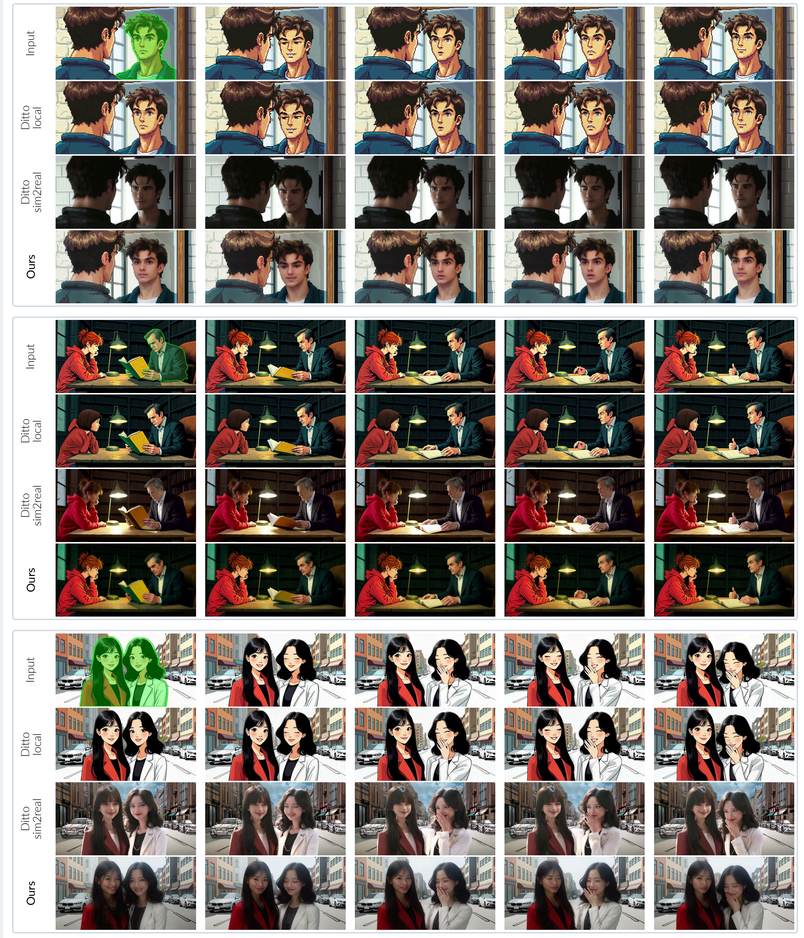


FFT (r = 32) DT-CWPT (r = 32) FFT (r = 64) DT-CWPT (r = 64)

FFT introduces **ringing artifacts**, while **wavelet** packets preserve **better** structure.

Instance-level Translation

Goal: Translate the highlighted instance into the real domain while preserving the style of all other regions.



DT-CWPT's spatial **locality** enables instance-level translation by confining edits to the selected region.

Results



Method	Control modules	Paired data	sKID ↓	sFID ↓	CLIP-IQA ↑	MS (%) ↑	ADE (m) ↓ 1s	ADE (m) ↓ 2s	ADE (m) ↓ 3s	ADE (m) ↓ Avg.	FDE (m) ↓
Input	–	–	1.97	48.89	0.367	98.58	0.520	2.008	4.484	2.337	5.223
CT	Required	Not required	2.42	51.98	0.429	97.84	0.516 (-0.9%)	1.999 (-0.5%)	4.444 (-0.9%)	2.319 (-0.8%)	5.173 (-1.0%)
Ditto	Required	Required	1.43	**33.99**	**0.496**	98.01	0.495 (-4.8%)	2.022 (+0.7%)	4.635 (+3.4%)	2.384 (+2.0%)	5.424 (+3.8%)
NR	Not required	Not required	1.45	37.65	0.450	98.30	0.532 (+2.2%)	2.079 (+3.5%)	4.630 (+3.2%)	2.413 (+3.3%)	5.389 (+3.2%)
Ours	Not required	Not required	**1.35**	36.58	0.495	**98.58**	**0.485** (-6.7%)	**1.898** (-5.5%)	**4.251** (-5.2%)	**2.211** (-5.4%)	**4.959** (-5.1%)

VLA Logs



Scene: The front-view camera image displays ... On the right side of the road, a sidewalk with a curb exists, and a traffic light pole stands with its bottom light **illuminated green**. A large palm tree ...
Intent: The ego's previous intent was to **decelerate significantly** from ... It should now **accelerate**, aiming to ... as the **traffic light is green** and there are no visible vehicles or pedestrians on the road.

Scene: The camera view presents ... On the far right, a vertical traffic light pole holds a traffic signal displaying a **solid red light at the top** and a **red downward arrow light illuminated below it** ...
Intent: The ego's previous intent was to **decelerate significantly** from ... The current scene displays a **solid red traffic light with a red downward arrow**, the ego should **decelerate** to come to a **complete stop at 0 m/s** and ...

Scene: The scene depicts ... A traffic light pole is positioned on the right side of the road, and its vertical signal head shows a **green light illuminated at the bottom**. A large palm tree ...
Intent: The ego vehicle's previous intent was to **decelerate** ... The current scene has a **green traffic light** and no visible obstacles, it should **accelerate**. A smooth acceleration of approximately **1.5 to 2 m/s^2** would be appropriate ...

Code, Paper & Contact

contact@kaiwenwang.cv
kit-mrt.github.io/Wavelet-Phase-Diffusion

GitHub arXiv Hugging Face