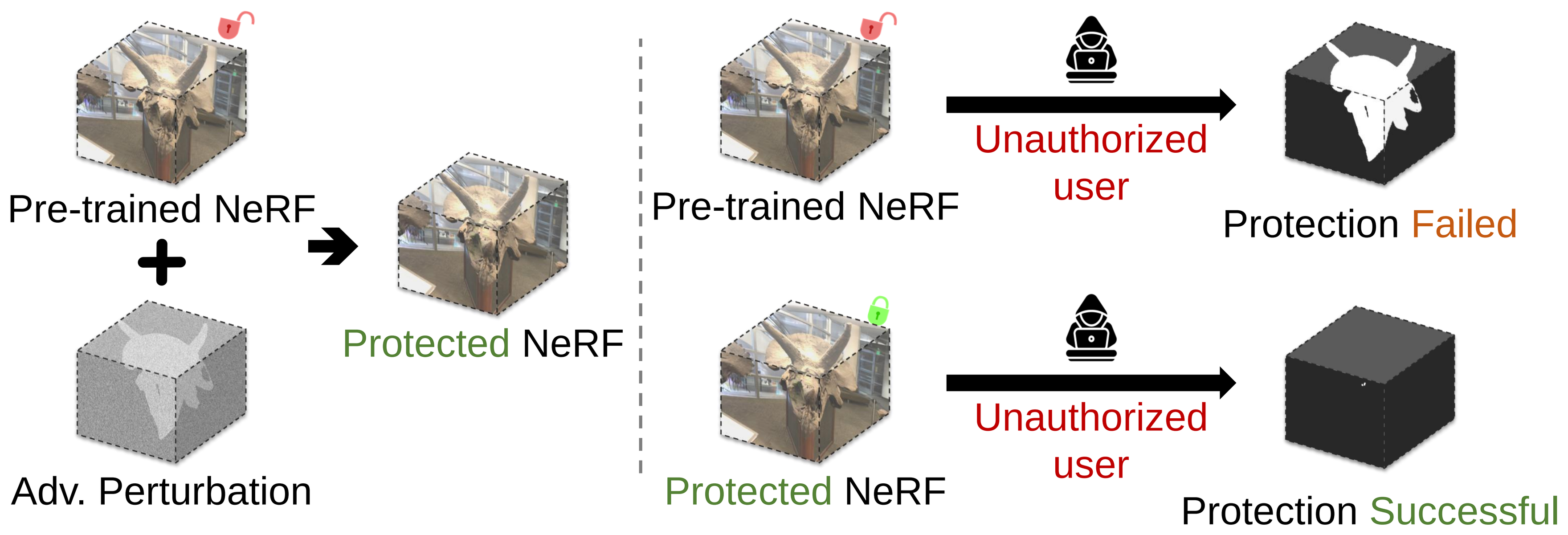
AegisRF: Adversarial Perturbations Guided with Sensitivity for Protecting Intellectual Property of Neural Radiance Fields

Woo Jae Kim, Kyu Beom Han, Yoonki Cho, Youngju Na,
Junsik Jung, Sooel Son, and Sung-Eui Yoon

KAIST

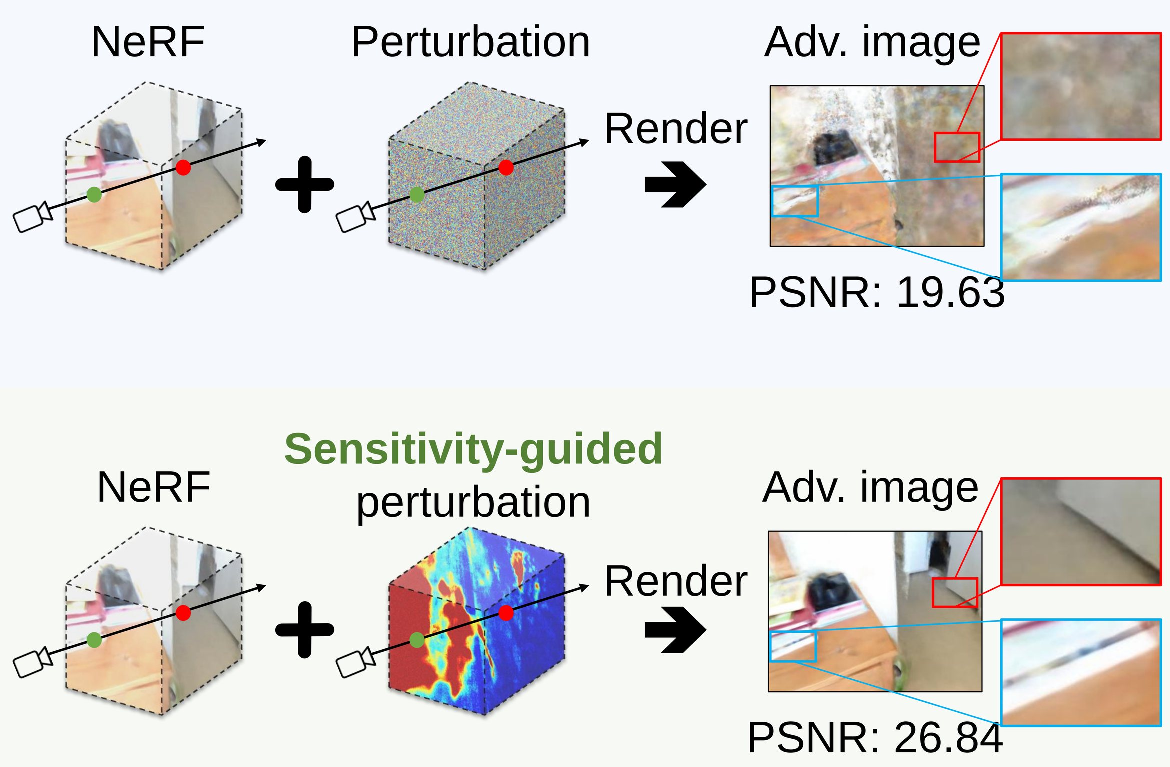
BMVC 2025

Intellectual property of NeRFs



- NeRFs are powerful 3D representation widely used in various downstream tasks
- Their unauthorized use can lead to loss in resources and revenues
- Goal: Prevent unauthorized downstream use of NeRF via adversarial perturbation

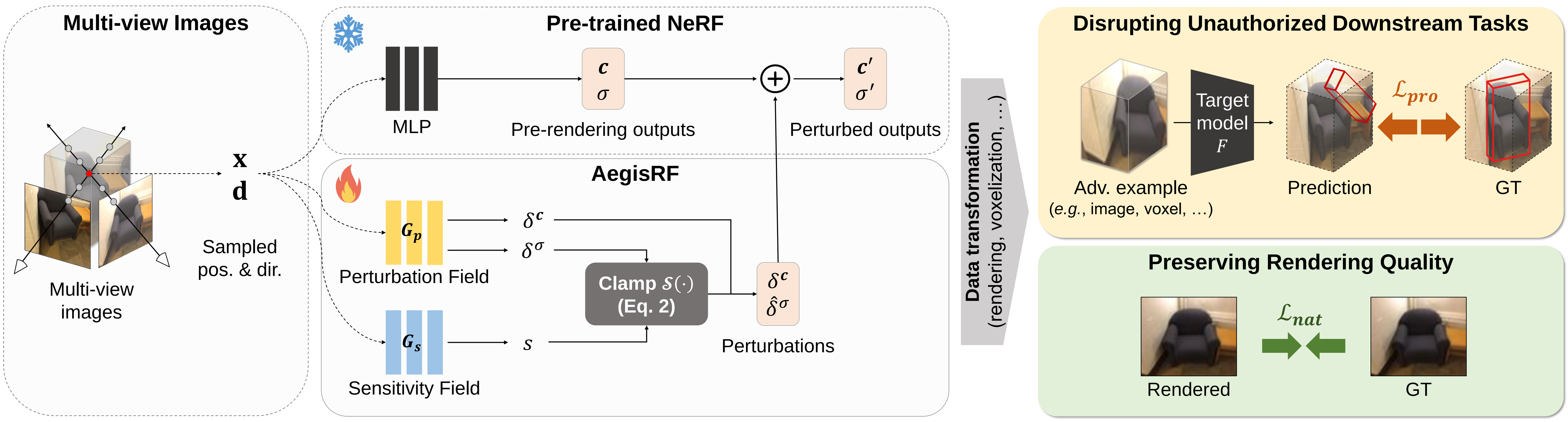
Challenge: Preserving rendering quality



Prior methods:
- Perturbs geometry indiscriminatively
- Poor rendering quality

Our approach:
- Perturbs geometry based on ***sensitivity*** to rendering quality
- Good rendering quality

AegisRF



Perturbation Field

- Goal: Generate perturbations on color and geometry pre-rendering outputs for each 3D point
- Approximated using a neural network ➔ generation of perturbation for any arbitrary point
- Applicable across diverse downstream models

Sensitivity Field

- Goal: Identify the sensitivity of each spatial region to rendering quality
- Predicts sensitivity score for each 3D point using a neural network
- Sensitivity score is used to clamp geometric perturbations

Model Training

Protection loss

$$\mathcal{L}_{\text{pro}} = -\mathcal{L}_F\big(F(x'), y_{gt}\big)$$

Naturalness loss

$$\mathcal{L}_{\text{nat}} = \sum_{\mathbf{r}\in\mathcal{R}} \| I'_n(\mathbf{r}) - I^{gt}_n(\mathbf{r}) \|_2^2$$

Evaluations

Quantitative Results

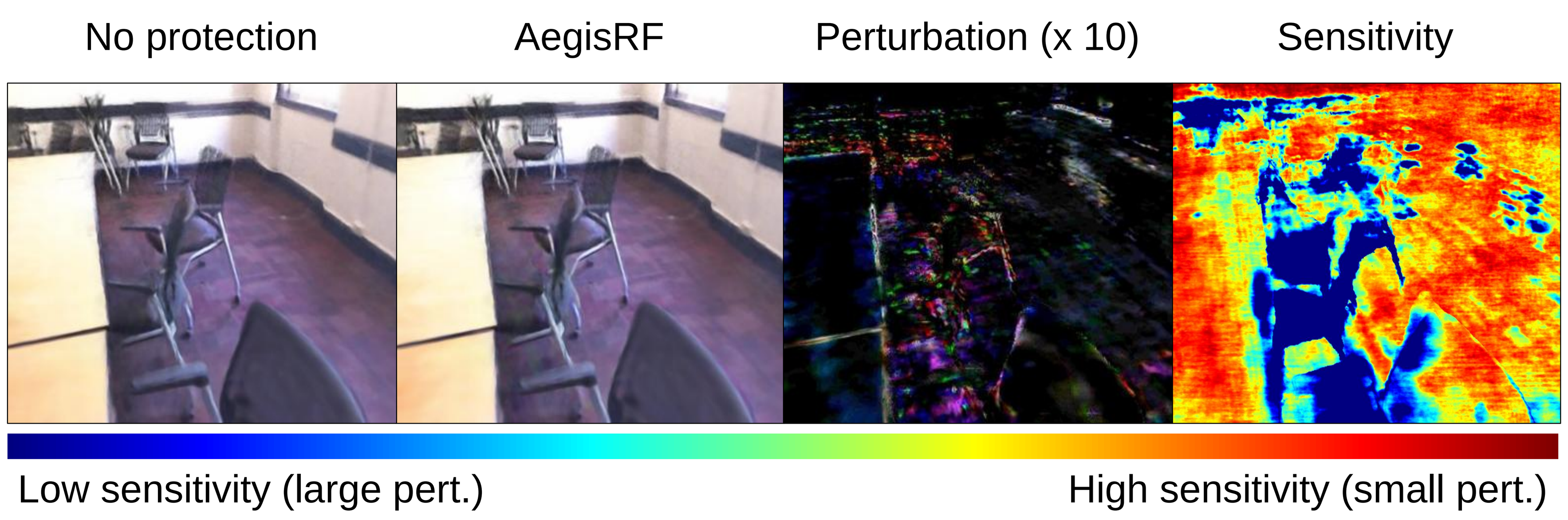
Multi-view classification

Method	Naturalness PSNR (↑)	Naturalness SSIM (↑)	Naturalness LPIPS (↓)	Disruption efficacy Acc. (%)
No protection	28.28	0.9205	0.0979	99.00
NeRFail-S	22.42	0.8360	0.1508	2.00
NeRFail	23.74	0.8751	0.1315	7.75
Adv-FT ($\lambda_{\text{pro}} = 0.0003$)	<u>25.57</u>	<u>0.8973</u>	<u>0.1372</u>	5.50
Adv-FT ($\lambda_{\text{pro}} = 0.003$)	23.67	0.8722	0.1987	2.13
Adv-FT ($\lambda_{\text{pro}} = 0.03$)	17.62	0.7442	0.5220	**0.13**
AegisRF (Ours)	**26.33**	**0.9042**	**0.1271**	<u>1.88</u>

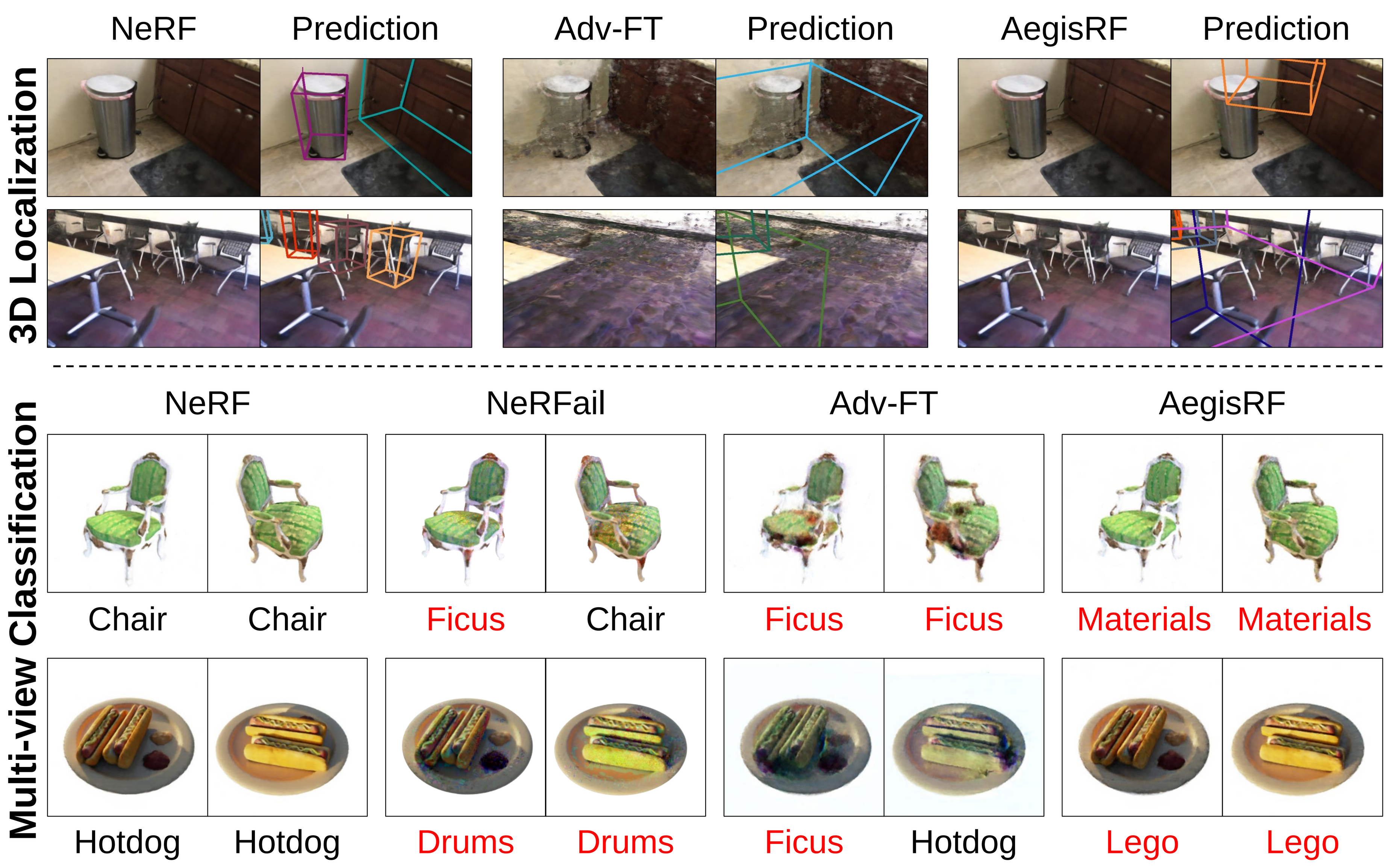
3D localization

Method	Naturalness PSNR (↑)	Naturalness SSIM (↑)	Naturalness LPIPS (↓)	Disruption efficacy Recall_{25} (%)	Recall_{50} (%)	AP_{25} (%)	AP_{50} (%)
No protection	23.98	0.7884	0.3186	95.44	66.63	59.94	44.35
Adv-FT ($\lambda_{\text{pro}} = 0.1$)	<u>22.37</u>	<u>0.7249</u>	<u>0.4289</u>	66.58	15.86	14.93	2.52
Adv-FT ($\lambda_{\text{pro}} = 1$)	18.44	0.6068	0.6013	50.82	6.88	3.83	0.22
Adv-FT ($\lambda_{\text{pro}} = 10$)	15.10	0.5264	0.7147	**43.43**	**2.25**	**1.60**	**0.01**
AegisRF (Ours)	**23.97**	**0.7687**	**0.3565**	<u>48.64</u>	<u>4.77</u>	<u>3.24</u>	<u>0.20</u>

Sensitivity Analysis



Qualitative Results



Webpage: sgvr.kaist.ac.kr/~wjkim/AegisRF
Codes: github.com/wkim97/AegisRF
Contact: woojae1123@gmail.com

Project page

Personal page