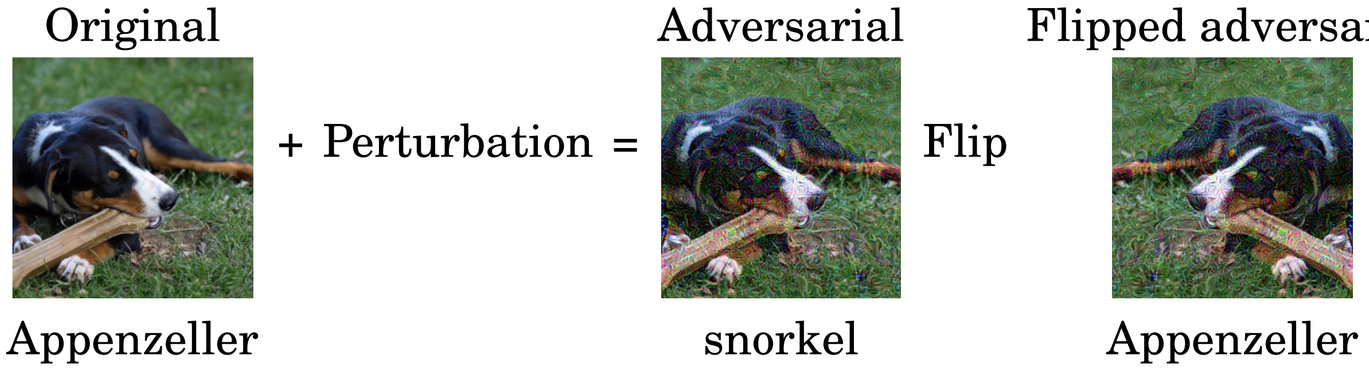
Secure Systems Group, Aalto University

Blerta Lindqvist

Symmetry Defense Against CNN Adversarial Perturbation Attacks

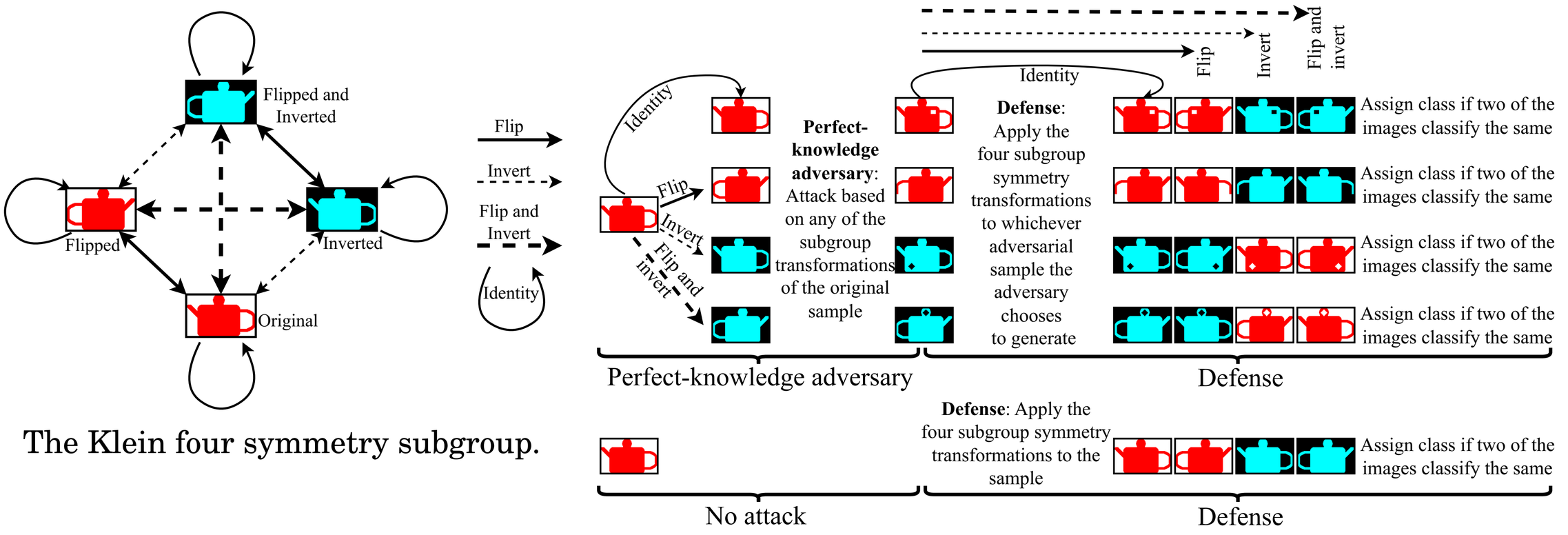
Problem: Adversarial perturbation causes image misclassification



Symmetry subgroup defense for white-box perfect-knowledge adversaries

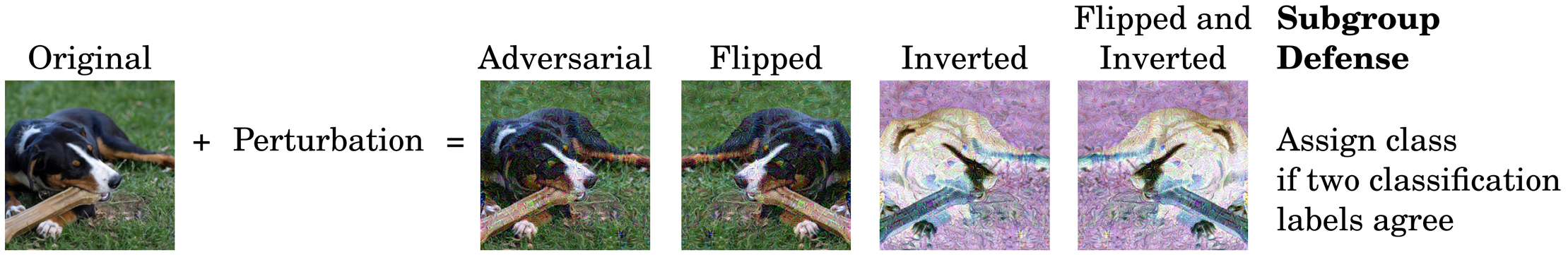
The proposed defense uses a symmetry subgroup that includes the flip symmetry and the pixel invert symmetry to defend against adversarial attacks with perfect-knowledge of the defense.

Symmetry subgroup **Symmetry subgroup defense**



The Klein four symmetry subgroup.

Results



Appenzeller	Perturbation	snorkel	**Appenzeller**	panda	**Appenzeller**	
76.19%	L_∞ PGD 4/255	0.00%	72.74%	33.09%	72.40%	70.59%
76.19%	L_∞ PGD 4/255	0.00%	72.87%	33.03%	72.52%	70.83%
76.20%	L_∞ PGD 4/255	0.00%	72.78%	34.56%	72.49%	70.71%
76.20%	L_∞ PGD 4/255	0.00%	72.98%	34.23%	72.49%	70.86%

Aalto University
School of Science

Contact: blerta.lindqvist@aalto.fi