敵対的サンプルを用いた白黒写真カラー化の防止
Uncolorable Examples: Preventing Illegitimate AI Colorization through Adversarial Attacks

Yuki Nii, Futa Waseda, Ching-Chun Chang, Isao Echizen, The University of Tokyo, National Institute of Informatics

Introduction

Motivation

Advances of AI-based colorization can lead to
Copyright infringements and **Fake media**

- Malicious users can colorize someone else's artwork and resell
  ex : A man in Japan was arrested for selling unauthorized colored version of the famous animation "Godzilla"
- Colorize historical images and alter artifacts making misleading media.
  ex : Color people's skin color differently

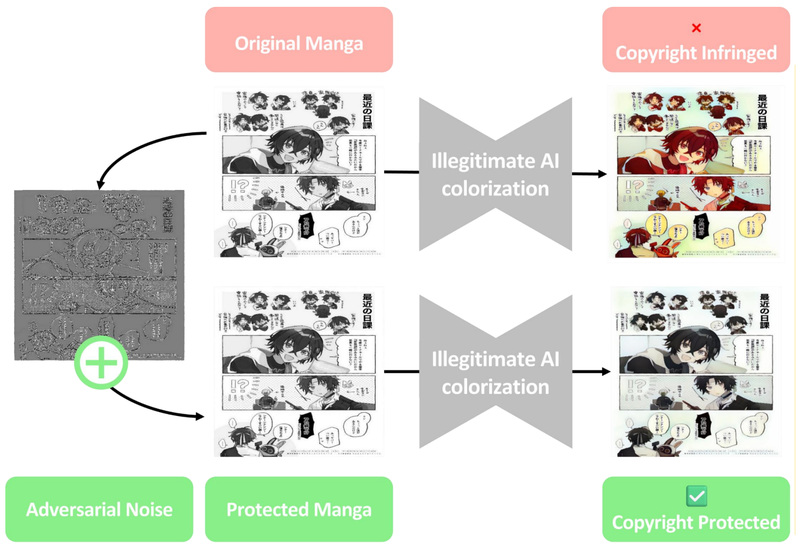
GOAL : **Neutralize** the colorization and make it **grayscale**.

Contribution

- Propose the first **defense baseline against unauthorized colorization**
- Ensure the perturbations are **robust, imperceptible** and **transferable**
- Evaluate the effects on architecturally different SOTA color models

[1]「モノクロ映画「ゴジラ」を無断でカラー化、海賊版ＤＶＤ販売の疑いで男逮捕…ＡＩ悪用し着色か」

Adversarial Attack



Proposing Framework

Propose **Mask Aware-Structural Invariant Attack (MA-SIA)**

Which utilizes **Adversarial Attacks** with

+ **Continuous Laplacian mask**
+ **Structural Invariant Attack (SIA)**

For Effective Suppression

Add human-invisible perturbations onto the input image to **mislead the colorization model** (**Adversarial Attack**)

To optimize the perturbations, we iteratively minimize the loss below to make the output grayscale

$$l_{CF} = \boldsymbol{Colorfulness}(\boldsymbol{G}(\boldsymbol{I} + \boldsymbol{\delta}))$$

$$\boldsymbol{Colorfulness}(\boldsymbol{CF}) = \sqrt{{\sigma^2}_{RG} + {\sigma^2}_{YB}} + 0.3\sqrt{{\mu^2}_{RG} + {\mu^2}_{YB}}$$

$$RG = |R - G|, \qquad YB = |0.5 \times (R + G) - B|$$

[3] D. Hasler and S. Suesstrunk, "Measuring colourfulness in natural images,"

Continuous Laplacian Mask

For Imperceptibility

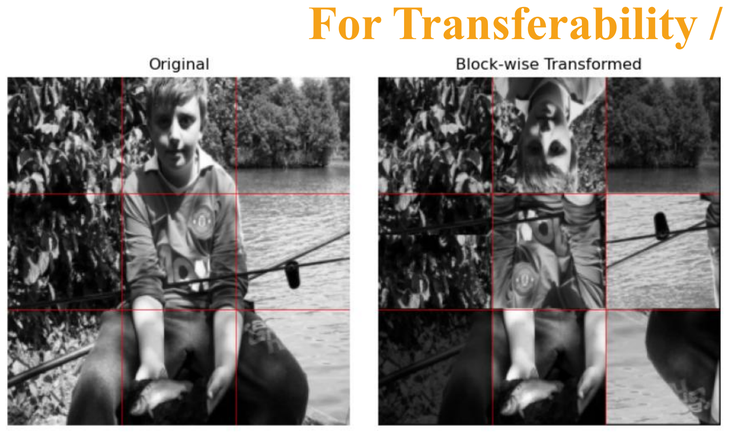
Laplacian Mask | Without Mask | With Mask

Suitable for **manga** with many plain area!

Make perturbations **concentrated within the masked areas.** Leverage the fact that **distortions are less likely to be perceived in edge regions.**

Structural Invariant Attack (SIA)

For Transferability / Robustness

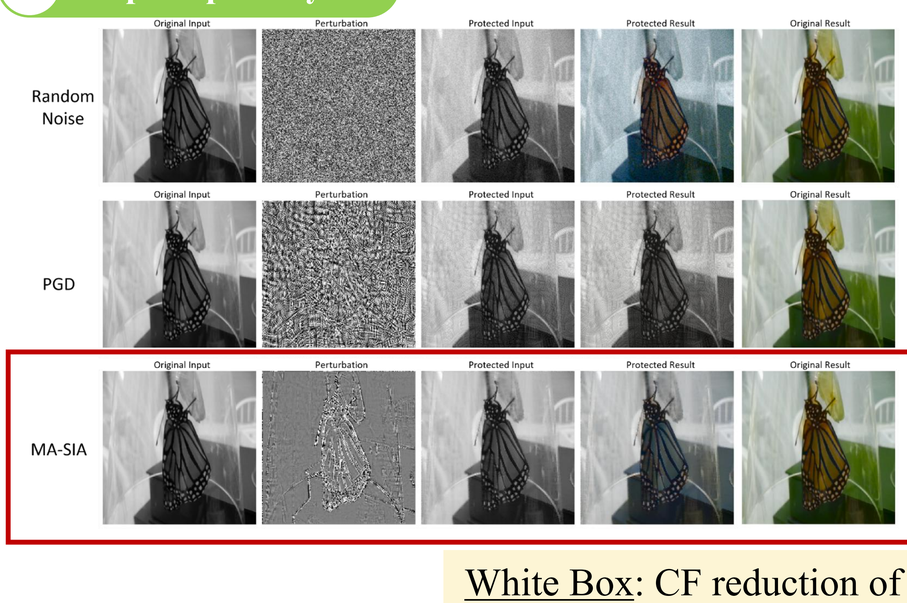


Make perturbation **robust to block-wise transformations.** Random vertical/ horizontal shifts and flips, 180°rotation, intensity scaling, DCT filtering, down/upsampling [2]

[2] X. Wang, Z. Zhang, and J. Zhang, "Structure invariant transformation for better adversarial transferability," 2023, arXiv:2309.14700.

Results

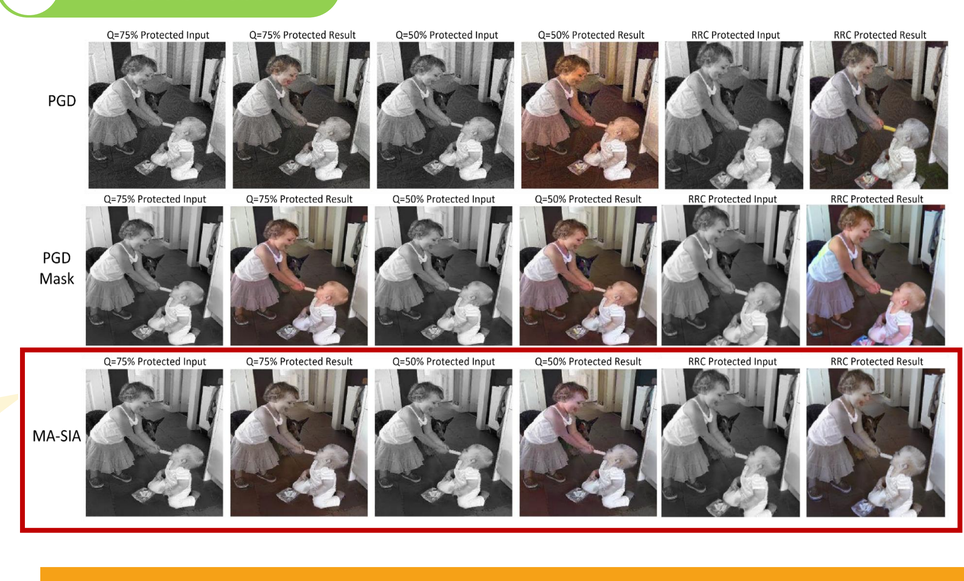
Imperceptibility



Comparing PGD to MA-SIA
CF reduction: **80~97%** → **78~82%**
PSNR (Input): **28dB** → **29.5dB**
SSIM (Input): **0.85** → **0.92**
→ Improvement of **imperceptibility** while keeping suppression

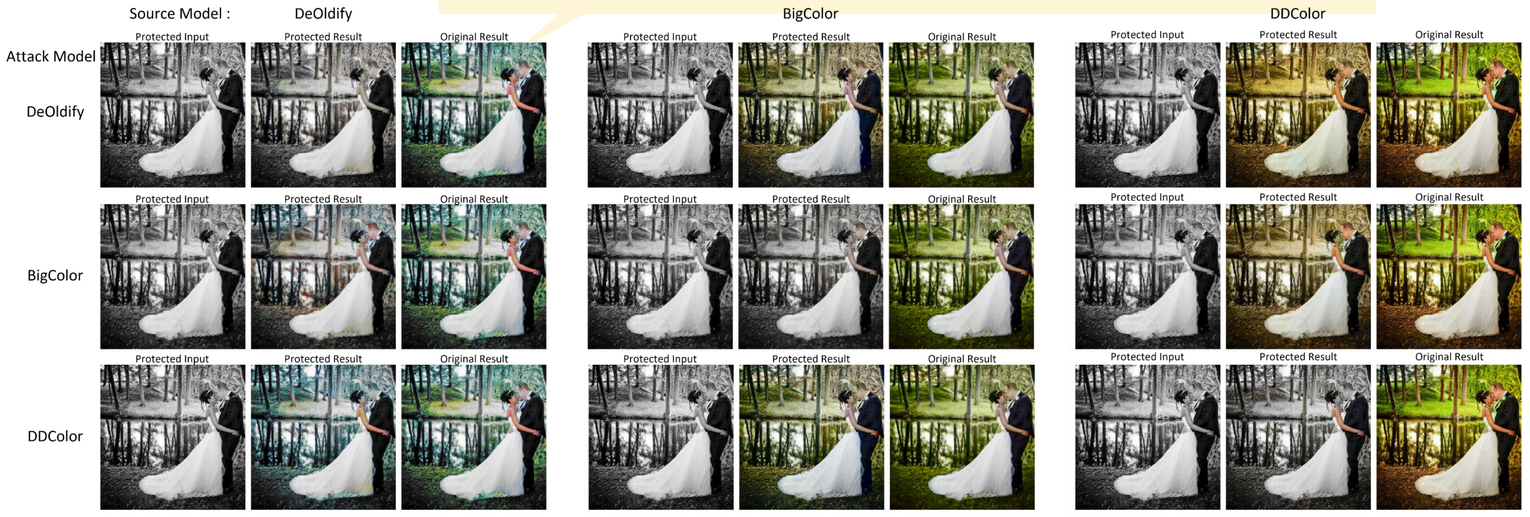
Even under JPEG compressions and random resized cropping (RRC) perturbations **stays effective**
→ Improvement of **robustness**

Robustness



Transferability

White Box: CF reduction of **MA-SIA** ~**80%**
Black Box: PGD-Mask 2~20% → **MA-SIA** **13~28%**
→ Improvement of **transferability**



Conclusion

- Proposed a baseline for preventing image colorization

Method	Effective	Robust	Imperceptible	Transferable
PGD	~90%	–	–	–
PGD Mask	~90%	–	✓	–
MA-SIA (Proposal)	~80%	✓	✓	✓

Future works

- Attack multi-modal media (ex. **Video**, sketch)
- Attack **user-hint based colorization models** such as the reference image / scribble / pallet based colorizations models

NII 越前研究室
国立情報学研究所
yuki_nii@nii.ac.jp