ALIGNING SUBJECTIVE AND OBJECTIVE ASSESSMENTS IN SUPER-RESOLUTION MODELS

Muhammad Hamza Zafar • Jon Y. Hardeberg

COLOURLAB NTNU, Norway

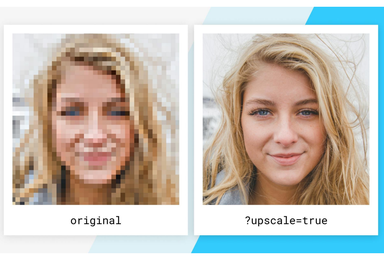
RESEARCH PROBLEM

Current SR evaluation relies heavily on objective metrics (PSNR, SSIM) but lacks human perspective validation.

- High **PSNR ≠ visually** pleasing results
- **Missing** human evaluation for SOTA models
- Bicubic paradox: high technical scores, poor visual quality
- Real-world applications **need perceptually satisfying results**

INTRODUCTION

- Super-resolution (SR) enhances image details
- **Objective metrics dominate** (e.g., PSNR, SSIM)
- **Need:** Incorporate human perceptual assessments



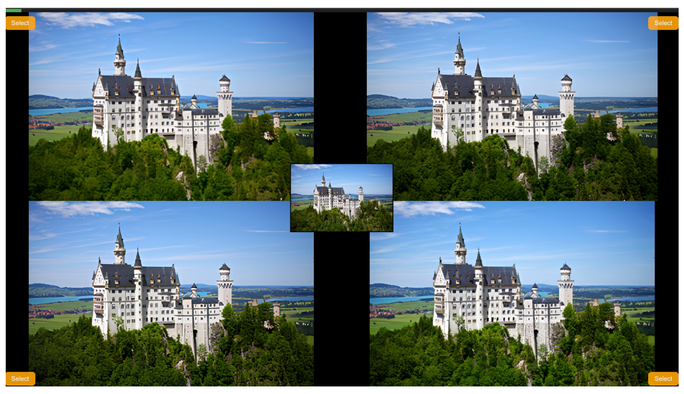
CONTACT & CODE

muhamz@stud.ntnu.no

https://github.com/hamzafer/super-resolution-color

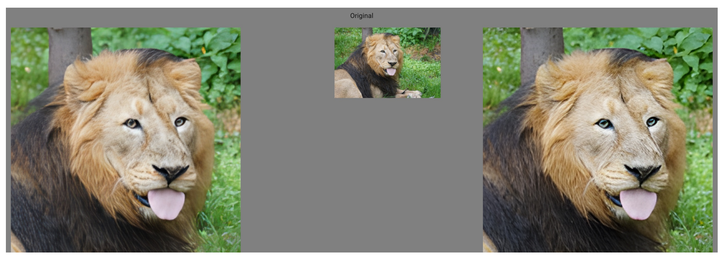
Project Page Contact

EXPERIMENT 1. ONLINE STUDY



EXPERIMENT SETUP

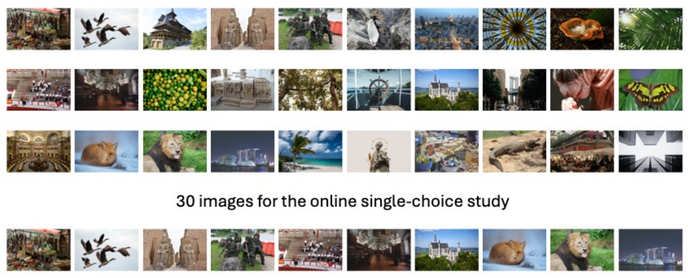
EXPERIMENT 2. CONTROLLED LAB



EXPERIMENT SETUP

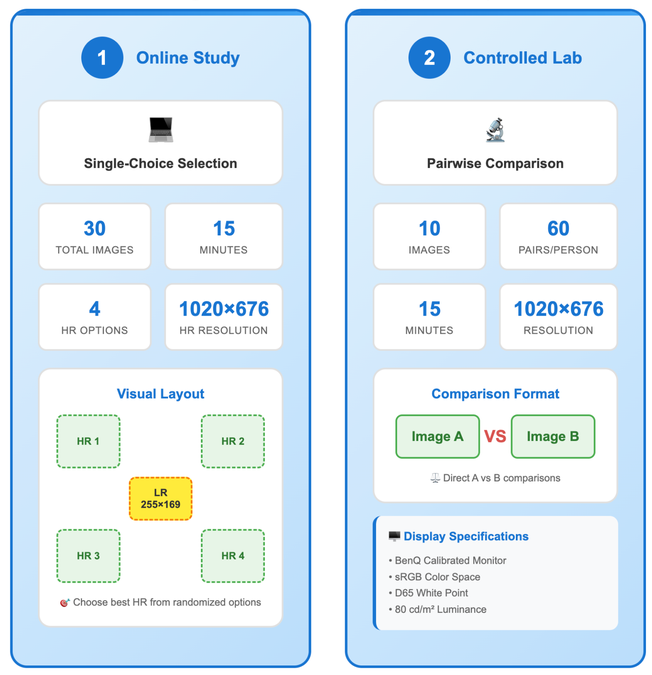
METHODOLOGY

- No additional preprocessing.
- Setup follows official **benchmark** protocol
- Used in both objective and subjective evaluations:
  - 30 **images** for the online single-choice study
  - 10 **images** for the controlled lab pairwise comparison
    - Subset of the above 30 images.



30 images for the online single-choice study

10 images for the controlled lab pairwise comparison



OBJECTIVE EVALUATION

- Metrics: **PSNR, SSIM, LPIPS, CLIPIQA** on 30 DIV2K Images
- ResShift leads in **4/4 metrics**
- BSRGAN shows second-highest PSNR
- RealESRGAN underperforms across most metrics
- SwinIR surprisingly lags in this comparison

Performance Heatmap

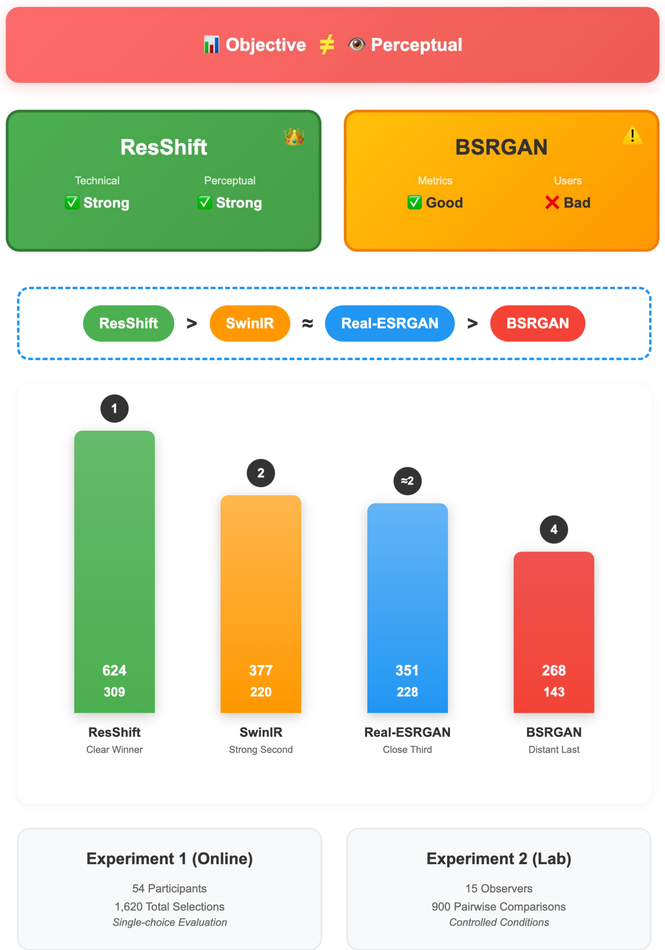
(Darker green shades indicate better performance within each metric)

Model	PSNR ↑	SSIM ↑	LPIPS ↓	CLIPIQA ↑
ResShift	24.65 ★	0.723 ★	0.136 ★	0.986 ★
BSRGAN	24.35	0.703	0.172	0.977
SwinIR	23.85	0.705	0.164	0.968
Real-ESRGAN	23.73	0.697	0.169	0.972

★ Best in metric — Performance scale

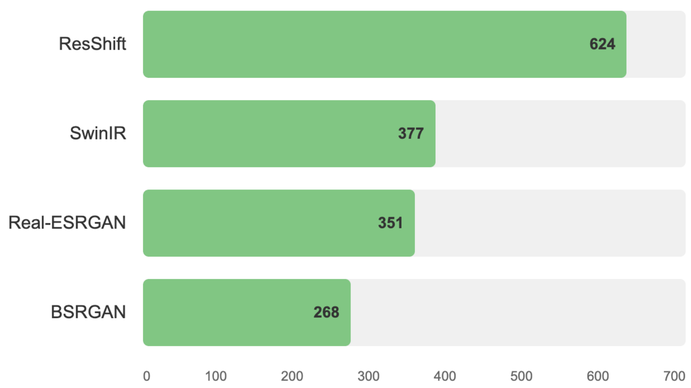
KEY FINDINGS

- Objective ≠ Perceptual always
- **ResShift** = technically & perceptually strong
- BSRGAN = metric-good, user-bad
- Hybrid evaluation is necessary for real-world quality
- Both studies show **identical** model rankings



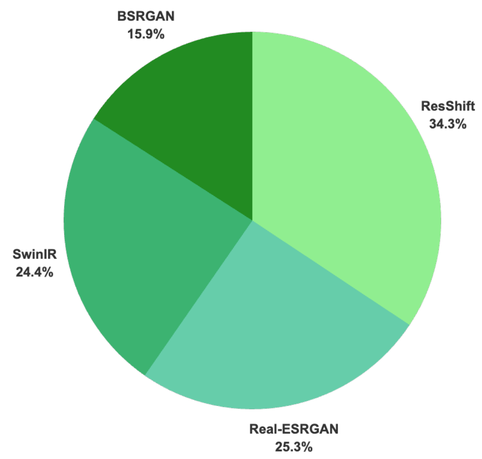
RESULTS

EXPERIMENT 1. ONLINE STUDY



MOST PICKED ALGORITHM

EXPERIMENT 2. CONTROLLED LAB



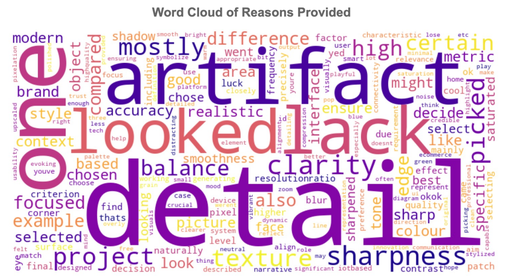
NUMBER OF SELECTION PER MODEL

LIMITATIONS

- **Limited Model Scope** — Only **4** models evaluated
- **Dataset Constraint** — DIV2K only (single benchmark dataset)
- **Small Lab Study** — Only **10** images (Experiment 2 dataset)
- **No Novel Metric** — Used existing evaluation methods only

CONCLUSION

Word Cloud of Reasons Provided



- **ResShift** sets a new standard in perceptual SR
- Need larger and more diverse datasets
- Explore joint metric-subjective learning for SR
- Push toward **human-centric model evaluation**

REFERENCES

1. IMAGE: https://www.imgix.com/blog/ai-powered-image-super-resolution
2. DATASET: Timofte, R., Gu, S., Wu, J., Van Gool, L., Zhang, L., Yang, M.H., Haris, M., et al.
3. NTIRE 2018 Challenge on Single Image Super-Resolution: Methods and Results.
4. Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW), June (2018)