

OR Paper Review

MaGIC: Multi-modality Guided Image Completion

Bruno Sánchez Gómez

April 24, 2025

Table of Contents

- 1 Introduction
- 2 MaGIC Overview
- 3 Critical Analysis

Definition

Image completion refers to the task of filling in missing regions within an image in a visually plausible way.

- **Applications:**

- **Inpainting:** Restoring damaged or missing parts of an image.
- **Outpainting:** Extending the boundaries of an image.
- **Editing:** Modifying images by adding or removing elements.

- **Approaches:**

- **Vanilla Image Completion:** Relies solely on existing image pixels around the masked region.
- **Guided Image Completion:** Uses external cues (e.g., text descriptions, edge maps, segmentation masks) for guidance.

Multi-modal Guided Image Completion (MaGIC) [1]

- **MaGIC:** A flexible framework for image completion guided by single or *arbitrary combinations* of modalities, such as:
 - Text
 - Canny Edge
 - Sketch
 - Segmentation
 - Depth
 - Pose
- **Architecture:** Based on pre-trained stable diffusion (SD) models with a U-Net denoiser.
- **Results:** Outperforms SOTA methods and generalizes well to various completion tasks.

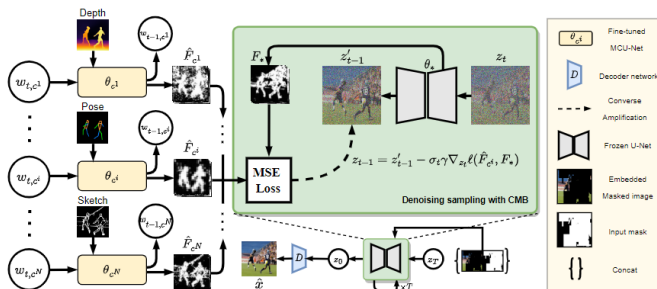
Table of Contents

1 Introduction

2 MaGIC Overview

3 Critical Analysis

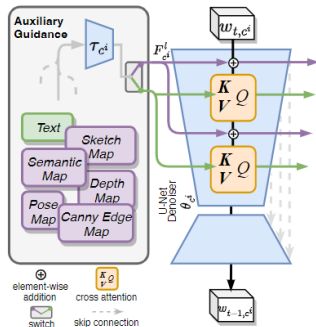
Components of MaGIC



- **Modality-specific Conditional U-Net (MCU-Net):** Injects single-modal guidance into a U-Net denoiser.
- **Consistent Modality Blending (CMB):** Training-free method to blend guidance from multiple pre-trained MCU-Nets, which enables easy addition of new modalities.

Modality-specific Conditional U-Net (MCU-Net)

- The encoding network τ_c is employed to extract multi-scale guidance signals, F_c^l .
- Each F_c^l is injected to the latent in MCU-Net to obtain modality-guided features.
- To leverage pre-trained SD, the U-Net denoiser is frozen. Only the encoding network τ_c is trained to extract guidance for the frozen denoiser.
- Achieves image completion under single-modality guidance.



Consistent Modality Blending (CMB)

- Uses a *converse amplification strategy* [2], which enables the intermediate feature maps F_* of a original U-Net to more closely approximate the MCU-Nets' guided feature maps \hat{F}_C

$$\begin{cases} z_{t-1} = z'_{t-1} - \sigma_t \gamma \nabla_{z_t} \ell(\hat{F}_C, F_*) \\ \ell(\hat{F}_C, F_*) = \frac{1}{N \cdot L} \sum_{c \in C} \sum_{l=1}^L \delta_c \left\| \hat{F}_c^l - F_*^l \right\|_2^2 \end{cases}$$

- **Properties:**

- It is *training-free*, as it operates on already trained MCU-Nets.
- Allows for *arbitrary combination* of available modalities.
- Straightforward *integration of new modalities*, by simply training a new MCU-Net for them. Avoids complex joint re-training.

Quantitative Results

Method	COCO		Places2		
	FID↓	PickScore↑ / %	FID↓	U-IDS↑ / %	P-IDS↑ / %
EC (Nazeri et al., 2019) ♠	76.64	23.14	25.08	12.89	2.86
CTSDG (Guo et al., 2021) ♠	97.05	24.03	42.81	0	0
ZITS (Dong et al., 2022) ♠	61.27	28.09	18.96	18.75	7.20
Our MCU-Net†	47.70±0.29	30.79±0.10	10.74±0.07	23.83±0.30	10.18±0.48
Our MCU-Net ♡	39.43±0.26	37.12±0.11	9.09±0.04	25.34±0.29	10.64±0.46
Our MCU-Net ♣	41.91±0.20	34.96±0.17	10.27±0.06	24.21±0.24	9.93±0.38
Our MCU-Net ♠	41.15±0.27	34.94±0.06	8.32±0.02	26.23±0.07	10.96±0.33

Table: Comparison of using single auxiliary modality as guidance for image completion. ♠: ground truth edge map as guidance, ♡: estimated depth map as guidance, ♣: segmentation map as guidance, ↑: the higher the better, ↓: the lower the better, †: completion without any guidance.

Qualitative Results

Image examples from the paper

Table of Contents

1 Introduction

2 MaGIC Overview

3 Critical Analysis

Why did MaGIC succeed?



Where did MaGIC fail?

- **Generalization:** MaGIC's framework can be applied to other image generation tasks, such as inpainting or super-resolution.
- **Modality Fusion:** The CMB method can be extended to fuse more complex modalities, such as audio or video.
- **Real-world Applications:** Potential applications in fields like medical imaging, autonomous driving, and augmented reality.

Thank you for your attention!

Any questions?

References I

-  Yongsheng Yu, Hao Wang, Tiejian Luo, Heng Fan, and Libo Zhang.
Magic: Multi-modality guided image completion.
arXiv preprint arXiv:2305.11818, 2023.
-  Prafulla Dhariwal and Alexander Quinn Nichol.
Diffusion models beat gans on image synthesis.
In *NeurIPS*, 2021.