# Bridging Global Context Interactions for High-Fidelity Image Completion


Chuanxia Zheng[1], Tat-Jen Cham[2], Jianfei Cai[1], Dinh Phung[1]

[1]Department of Data Science & AI, Monash University [2]School of Computer Science and Engineering, NTU

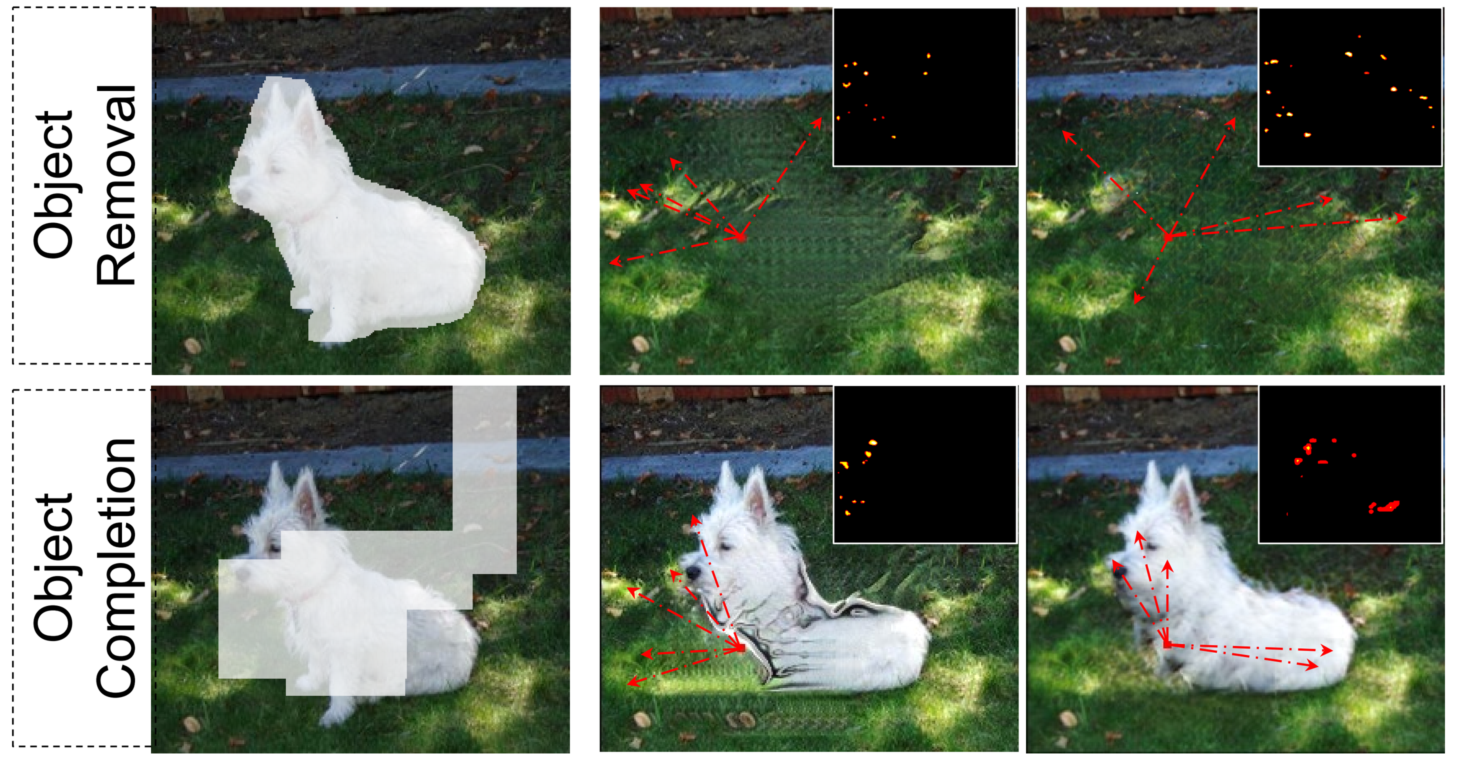

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

## Motivation

**Goal:** semantic image completion, *not* purely for removing objects

**Challenge:**

1. CNN *gradually* affected by neighboring pixels
2. Pixel-level attention requires *expensive* costs

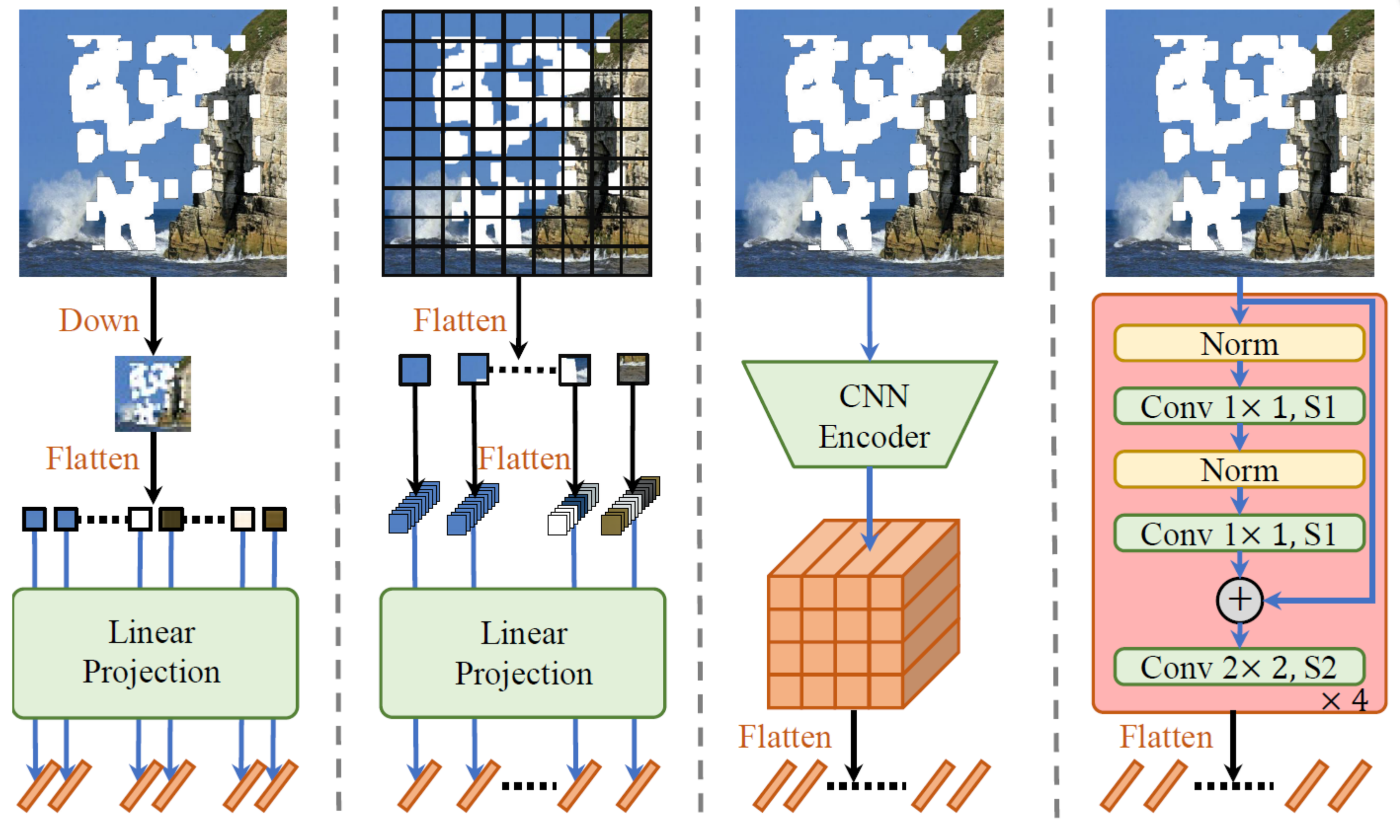



Masked Input HiFill[51]$_{CVPR'2020}$ TFill

How to correctly model the disconnected context?

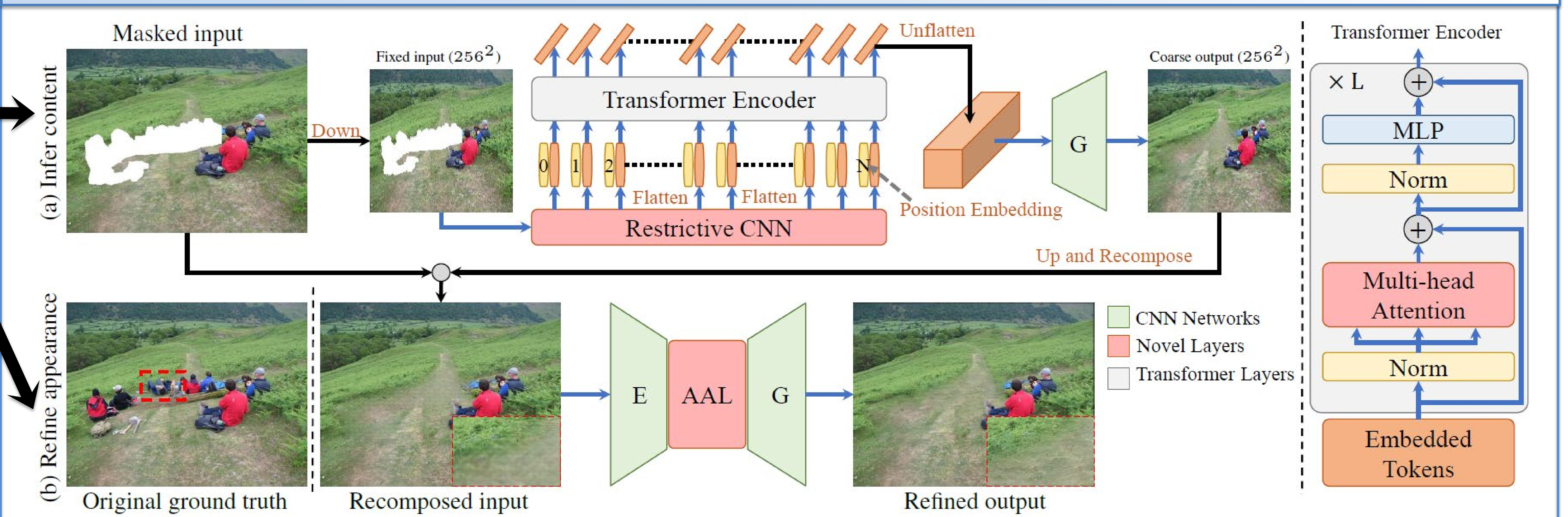
## Key Insights

1. Propose a transformer-based framework with the **restrictive CNN** to correctly model the **global content** in each attention layer
2. Design a novel Attention-Aware Layer to exploit **global appearance information**




iGPT[8] ViT[11] VQGAN[12] Ours

## Pipeline: Two-stage Image Completion




### TFill-*Coarse*: Global Content Inference

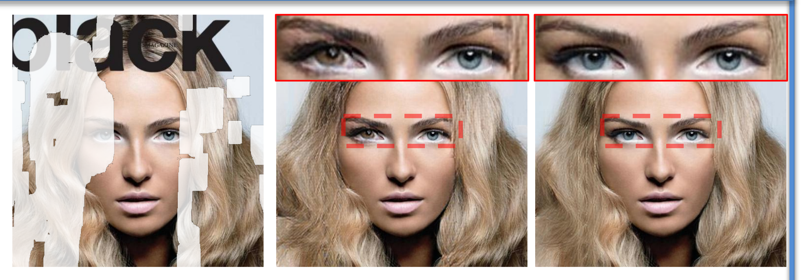
**Restrictive CNN**: embed patch visible context

**Weighted Self-Attention Layer**: bias visible token

**CNN-based Decoder**: generate tokens in parallel

### TFill-*Refined*: Global Appearance Refinement

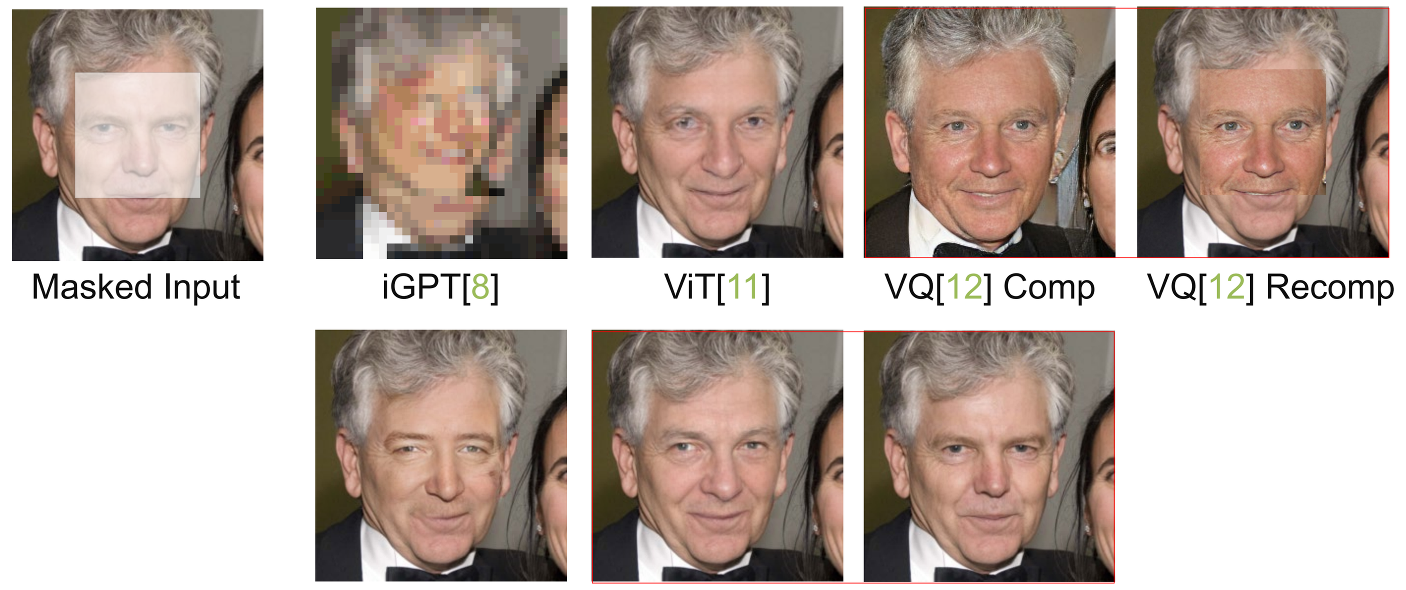
**AAL**: copy global context from both encoded and decoded features



Masked Input TFill-*Coarse* TFill-*Refined*

## Analysis: Architecture of TFill

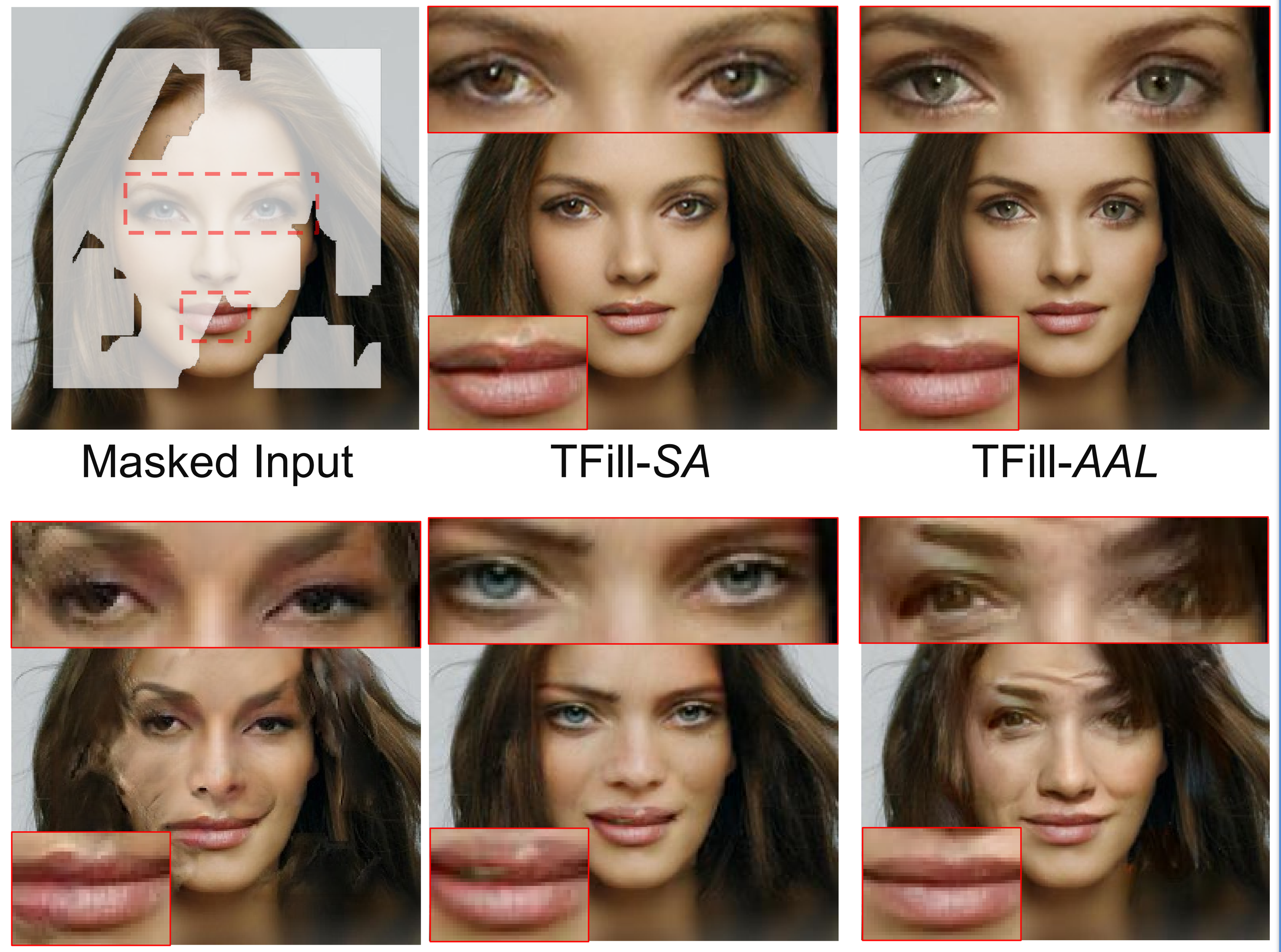
| Method | CelebA-HQ | | FFHQ | |
|---|---|---|---|---|
| | LPIPS↓ | FID↓ | LPIPS↓ | FID↓ |
| CA [52]$_{CVPR'2018}$ | 0.104 | 9.53 | 0.127 | 8.78 |
| PIC [60]$_{CVPR'2019}$ | 0.061 | 6.43 | 0.068 | 4.61 |
| MEDFE [29]$_{ECCV'2020}$ | 0.067 | 7.01 | - | - |
| A Traditional *Conv* | 0.060 | 6.29 | 0.066 | 4.12 |
| B + Attention in G | 0.059 | 6.34 | 0.064 | 4.01 |
| C + Restrictive *Conv* | 0.056 | 4.68 | 0.060 | 3.87 |
| D + Transformer | 0.051 | 4.02 | 0.057 | 3.66 |
| E + Masked Attention | 0.050 | 3.92 | 0.057 | 3.63 |
| F + Refine Network | **0.048** | **3.86** | **0.053** | **3.50** |



Masked Input iGPT[8] ViT[11] VQ[12] Comp VQ[12] Recomp

ICT[47] w/ Refine TFill-*Coarse* TFill-*Refine*

## Analysis: Attention-Aware Layer

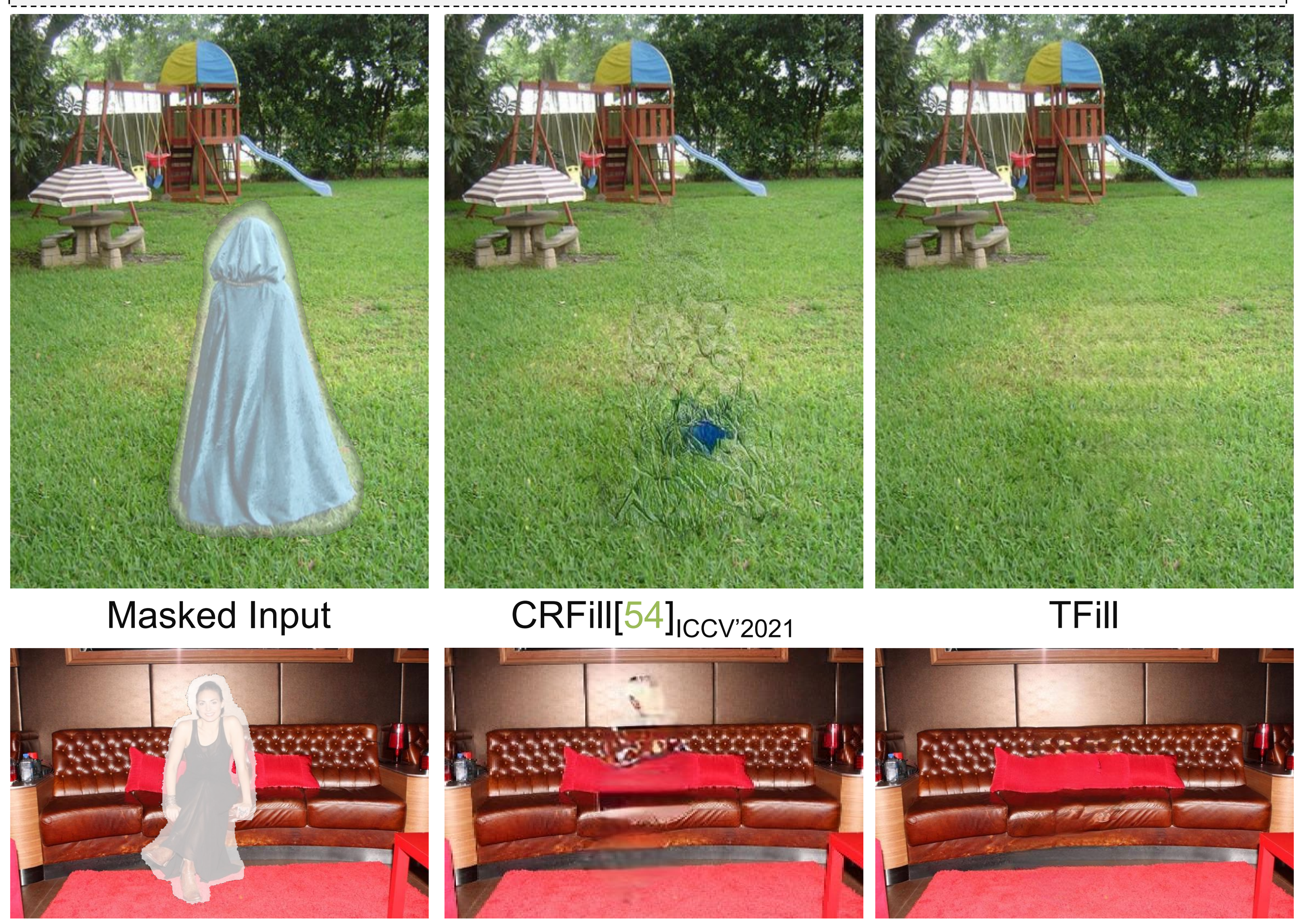


Masked Input TFill-*SA* TFill-*AAL*

CA[52]$_{CVPR'2018}$ PIC[60]$_{CVPR'2019}$ CRFill[54]$_{ICCV'2021}$

## Overall Comparison with Other Methods

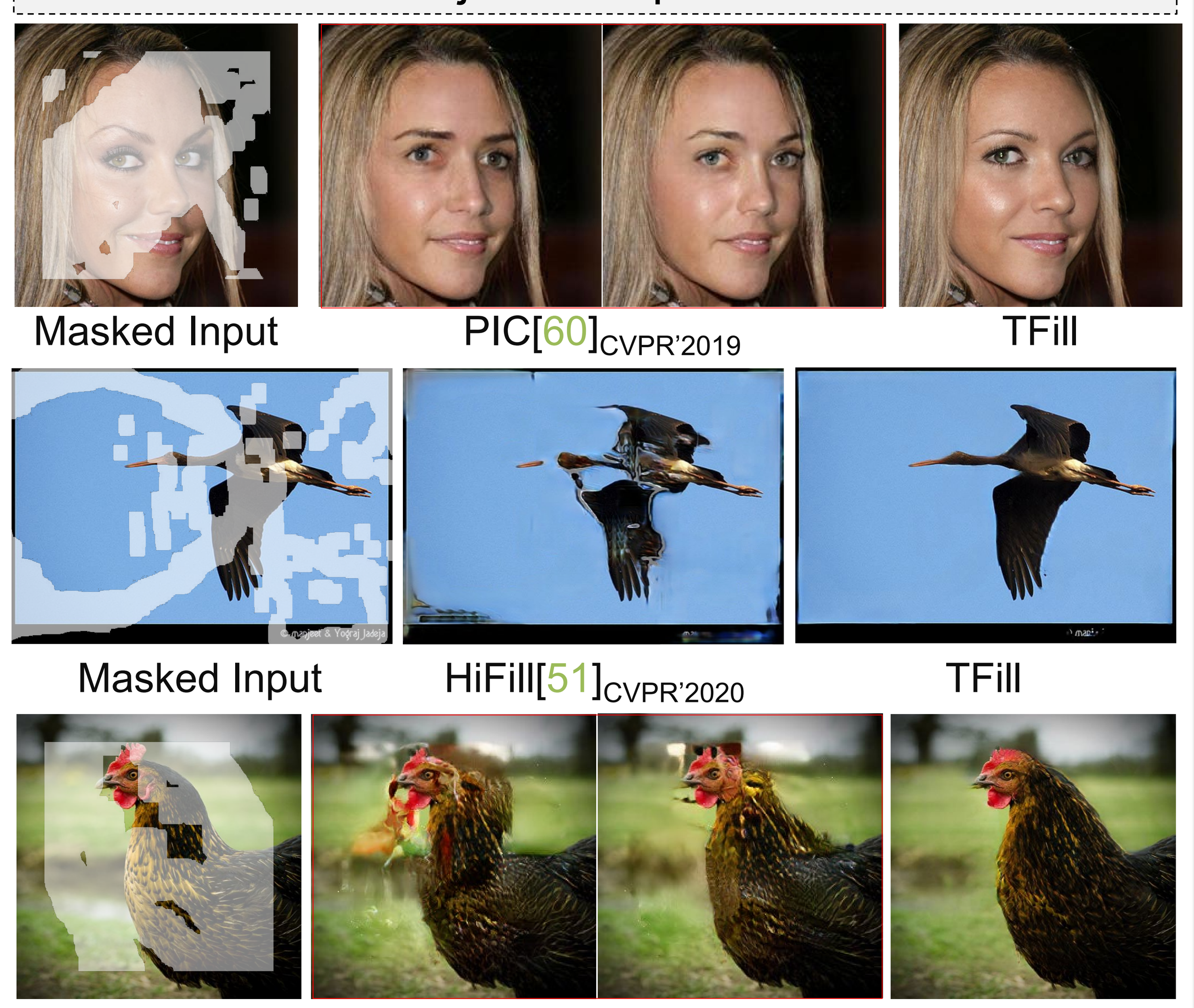
### Object Removal



Masked Input CRFill[54]$_{ICCV'2021}$ TFill

Masked Input DSI[37]$_{CVPR'2021}$ TFill

### Object Completion



Masked Input PIC[60]$_{CVPR'2019}$ TFill

Masked Input HiFill[51]$_{CVPR'2020}$ TFill

Masked Input ICT[47]$_{ICCV'2021}$ TFill