RGB↔X: Image decomposition and synthesis using material- and lighting-aware diffusion models

For GAMES Webinar 317

Observation

- Traditional rendering
  - + Precise
  - + Photo-realistic
  - - Requires full scene description

- Diffusion models
  - + Simple to use
  - + Confuse the real from the fake
  - - Hard for precise control

Idea

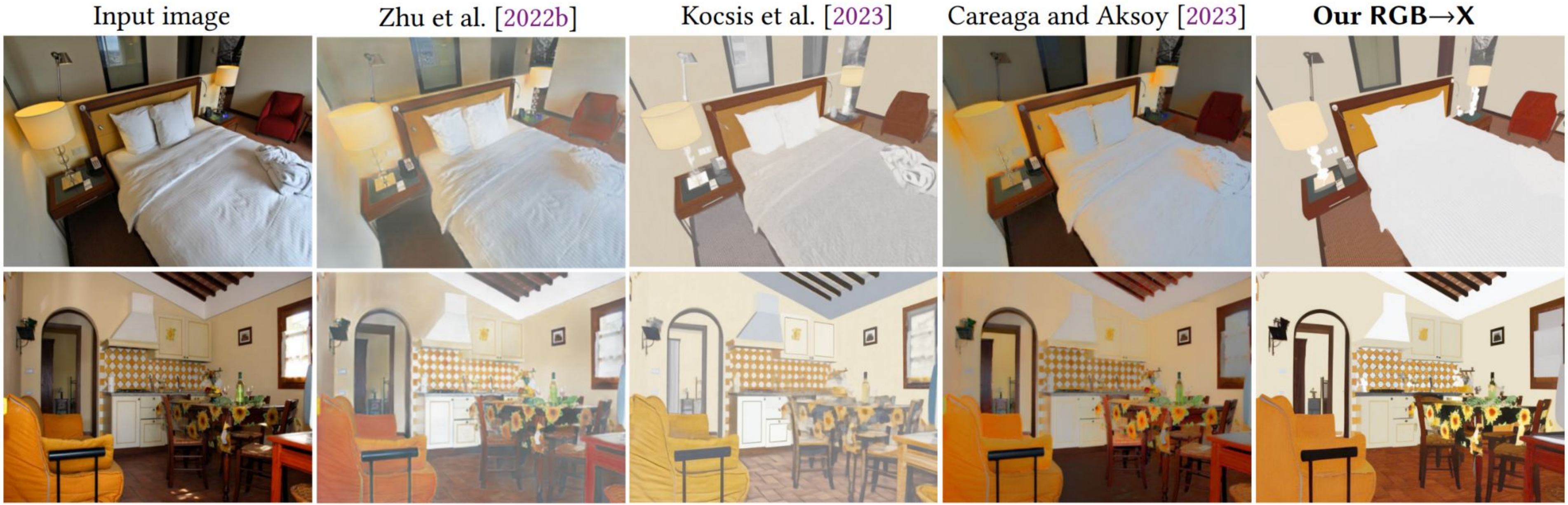
- We aim to explore a middle ground
  - specify only certain appearance properties, and
  - give freedom to the model to hallucinate a plausible version of the rest
- X: intrinsic channels (G-buffers)
- X -> RGB: synthesizing an image from a given description
- RGB -> X: decomposing an image into intrinsic channels

Background

- RGB->X: estimating per-pixel information from image
  - We denote these intrinsic channels (or, G-buffers) as X
- This problem is under-constrained and ambiguous
  - “Wooden floor with shadows and reflections on it”

Background

- Recent work show improved estimation on X based on diffusion models



Goal

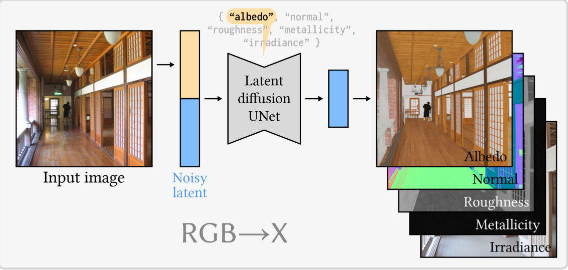
- Explore the connections between
  - diffusion models, rendering, and intrinsic channel estimation
- Focus on two problems
  - RGB->X: intrinsic channels estimation and
  - X->RGB: image synthesis conditioned on intrinsic channels

RGB->X

- Fine-tuned from pre-trained Stable Diffusion (latent diffusion model)
- Key idea:
  - repurpose the input text prompt as a “switch” to control the output,
  - produce a single intrinsic channel at a time
- Two benefits:
  - Enable usage of a mix of heterogeneous datasets, which differ in the available channels
    - For example, a dataset with only albedo channel available can still be employed to train our model
    - Massively enlarged the training datasets available to us.
  - Avoid handling multiple output channels
    - Which is proven to make the training more challenging

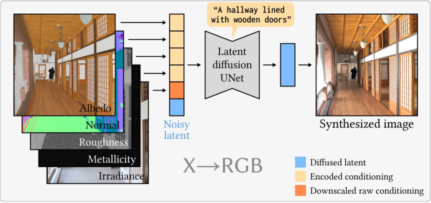
Table 1. We combine four heterogeneous datasets (ours in bold), each providing a subset of the channels we need for training. For each dataset we mark channels as available (✓), unavailable (✗), or available but not fully reliable (✓). We also include representative images from the datasets. ImageDecomp is an RGB-only dataset for which we estimated the intrinsic channels using our RGB→X model.

Dataset	Size	Albedo	Normal	Roughness	Metallic.	Irrad.
InteriorVerse	50,097	✓	✓	✓	✓	✗
Hypersim	73,819	✓	✓	✗	✗	✓
Evermotion	17,000	✓	✓	✓	✓	✗
ImageDecomp	50,000	✓	✓	✓	✓	✓

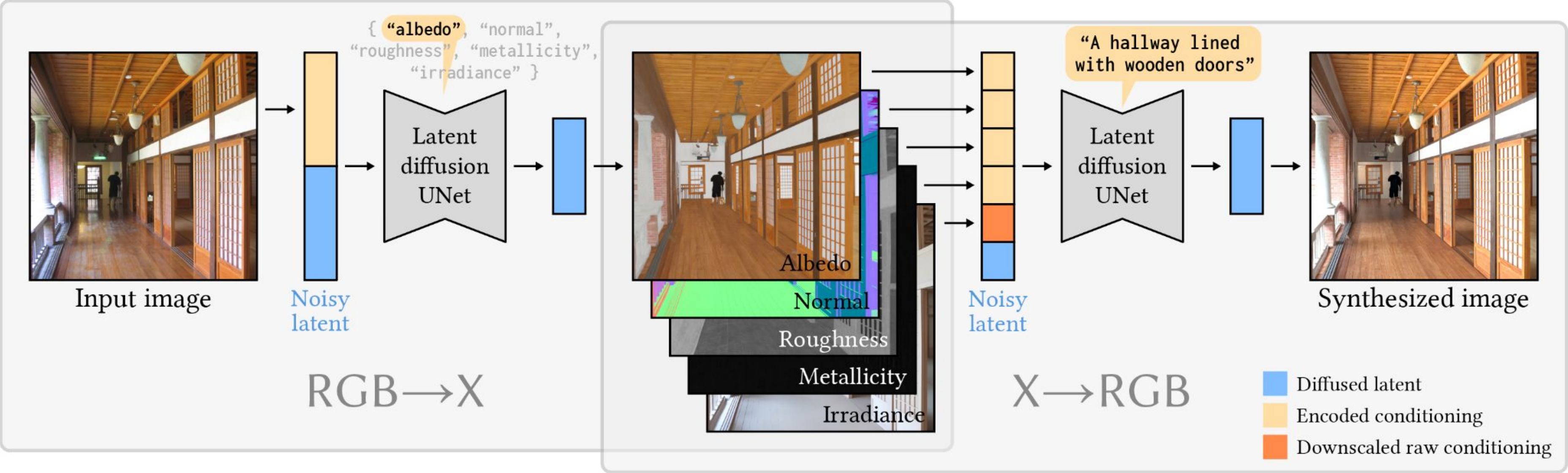


X->RGB

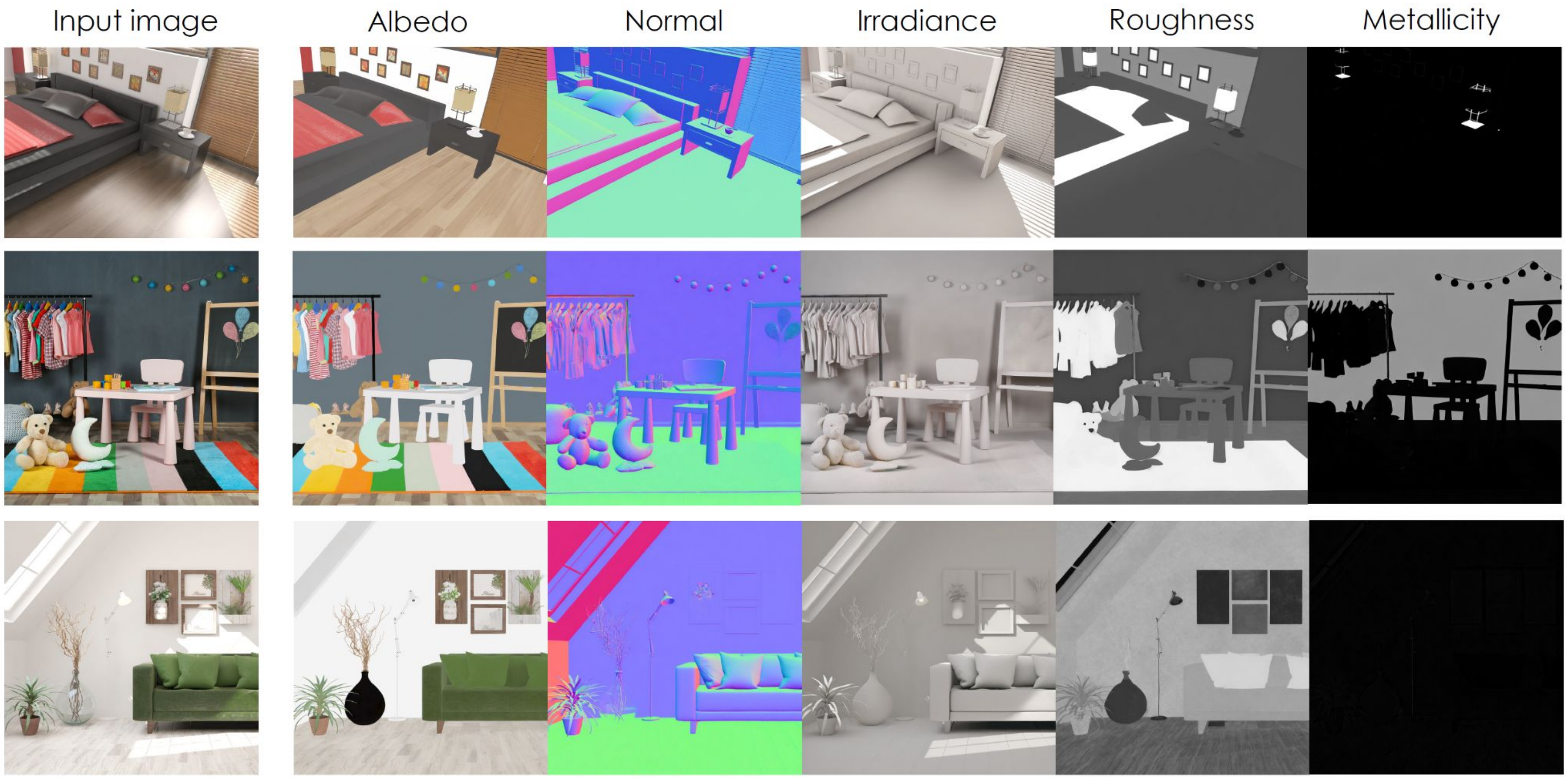
- Fine-tuned from pre-trained Stable Diffusion (latent diffusion model)
- Key idea:
  - A channel drop-out strategy: randomly drop conditioned channels during training.
    - For example, drop albedo channel with a probability of 0.3
  - Jointly train a conditional and unconditional diffusion model
- Two benefits
  - Again, enable usage of a mix of heterogeneous datasets, which differ in the available channels
  - Enable image generation with any subset of conditions
    - For example, providing no albedo or no lighting will result in the model generating plausible images, using its prior to compensate for the missing information



Full pipeline

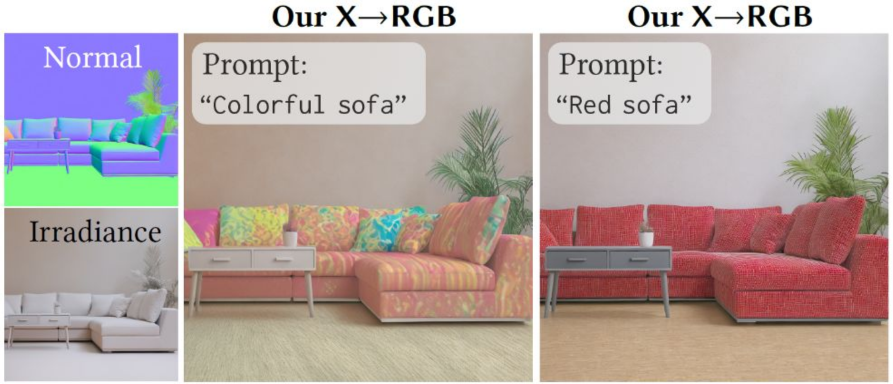


Results: RGB->X

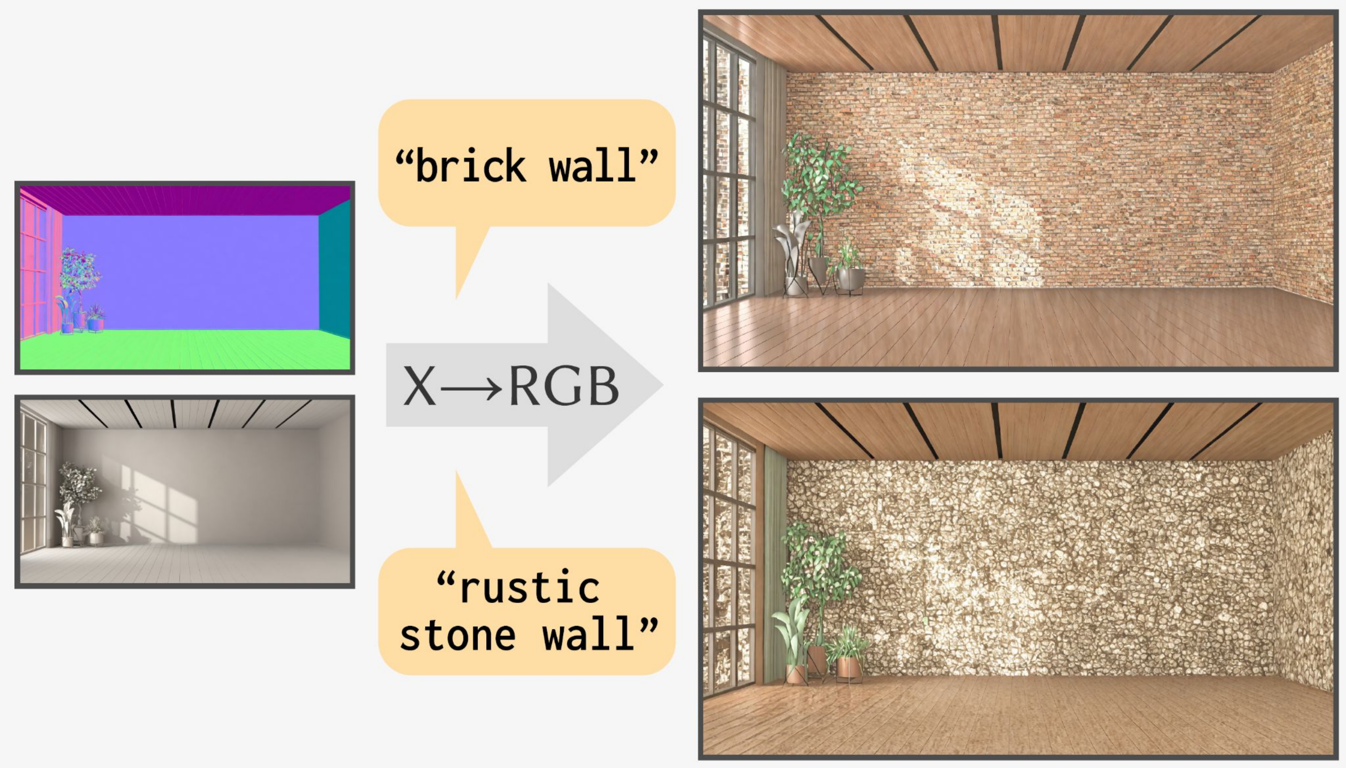


Results: X->RGB

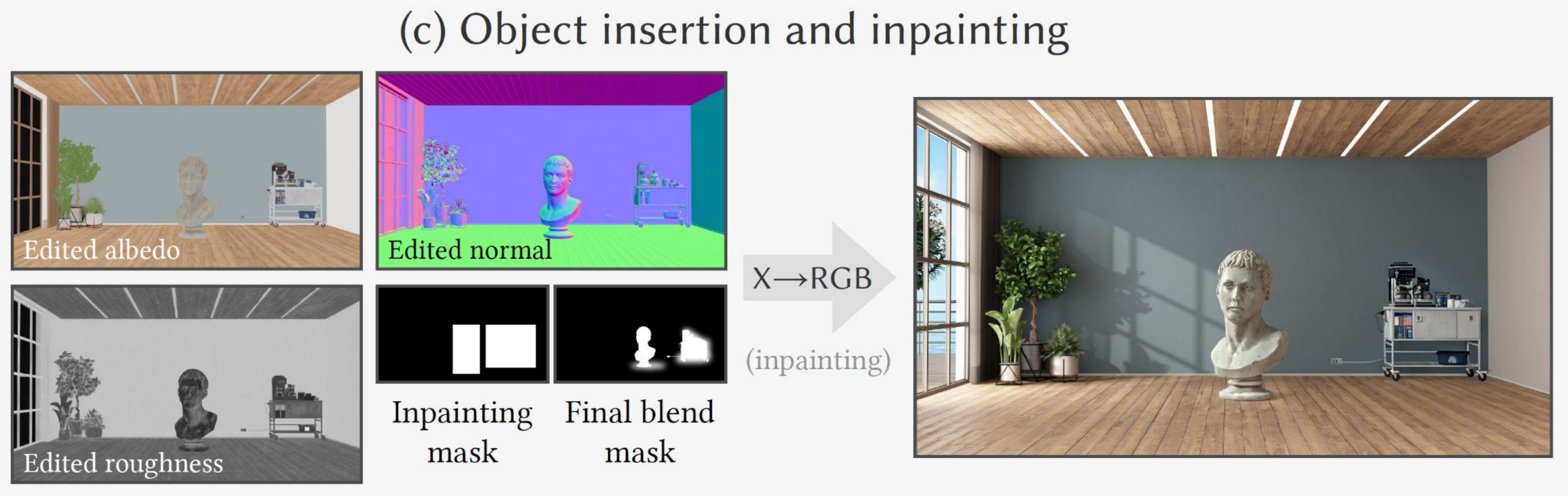
Results: X->RGB



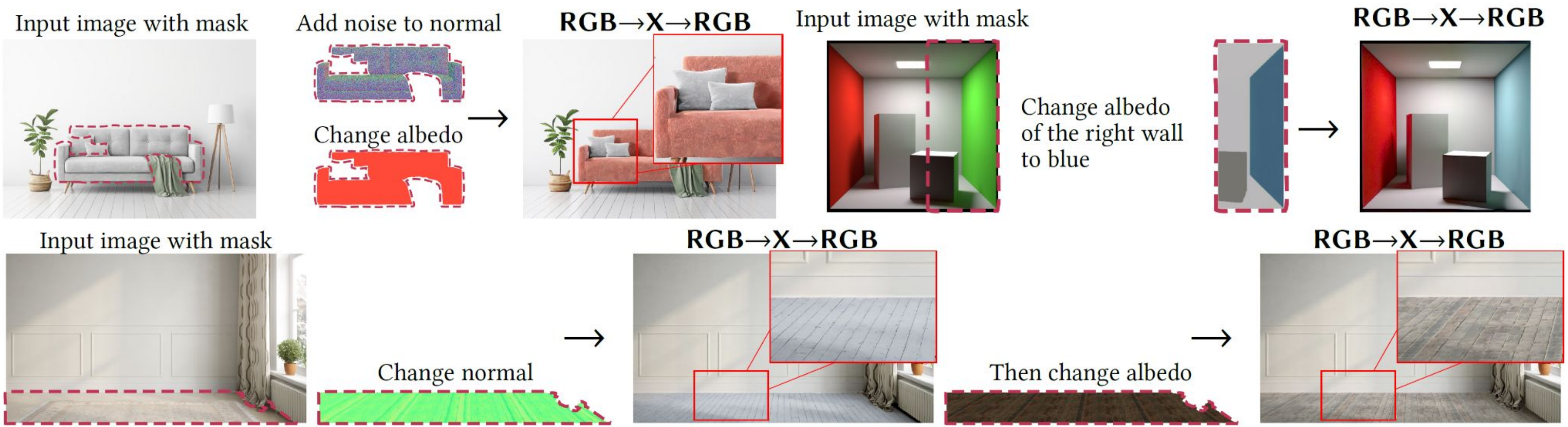
Results: X->RGB



Results: X->RGB



Results: RGB->X->RGB



Applications

- RGB->X
  - Estimation of intrinsic channels (albedo estimator, normal estimator, …)
- X->RGB
  - Fast previews of renderings for 3D software
- RGB->X->RGB
  - Material replacement, object insertion, relighting, …