ICML 2019

Compositional GAN: Learning Image-Conditional Binary Composition

Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, Trevor Darrell

University of California, Berkeley

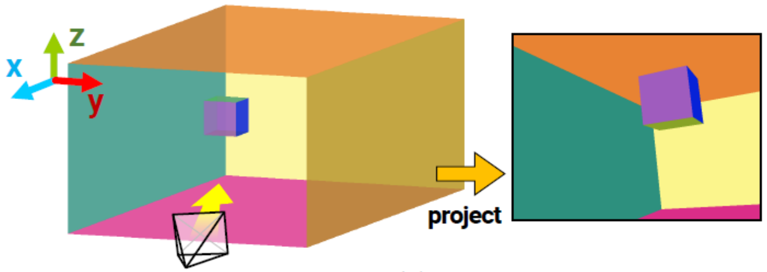
Revised after International Conference on Learning Representations (ICLR 2019)

Archived on Mar 28, 2019

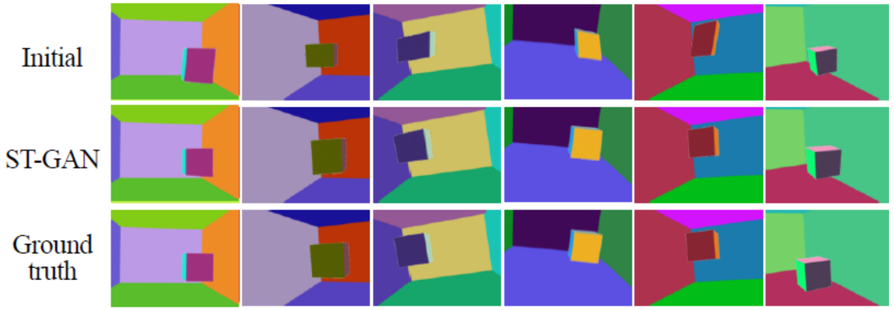
Slides compiled by Mengzhou Li

Introduction

- Modeling compositionality in natural images is a challenging problem due to complex interactions among different objects,
  - Relative scaling,
  - Spatial layout,
  - Occlusion,
  - Viewpoint transformation
- Recent work using spatial transformer networks (STN) within a GAN framework (ST-GAN) decomposes this problem by operating in a geometric warp parameter space to find a geometric modification for a foreground object.

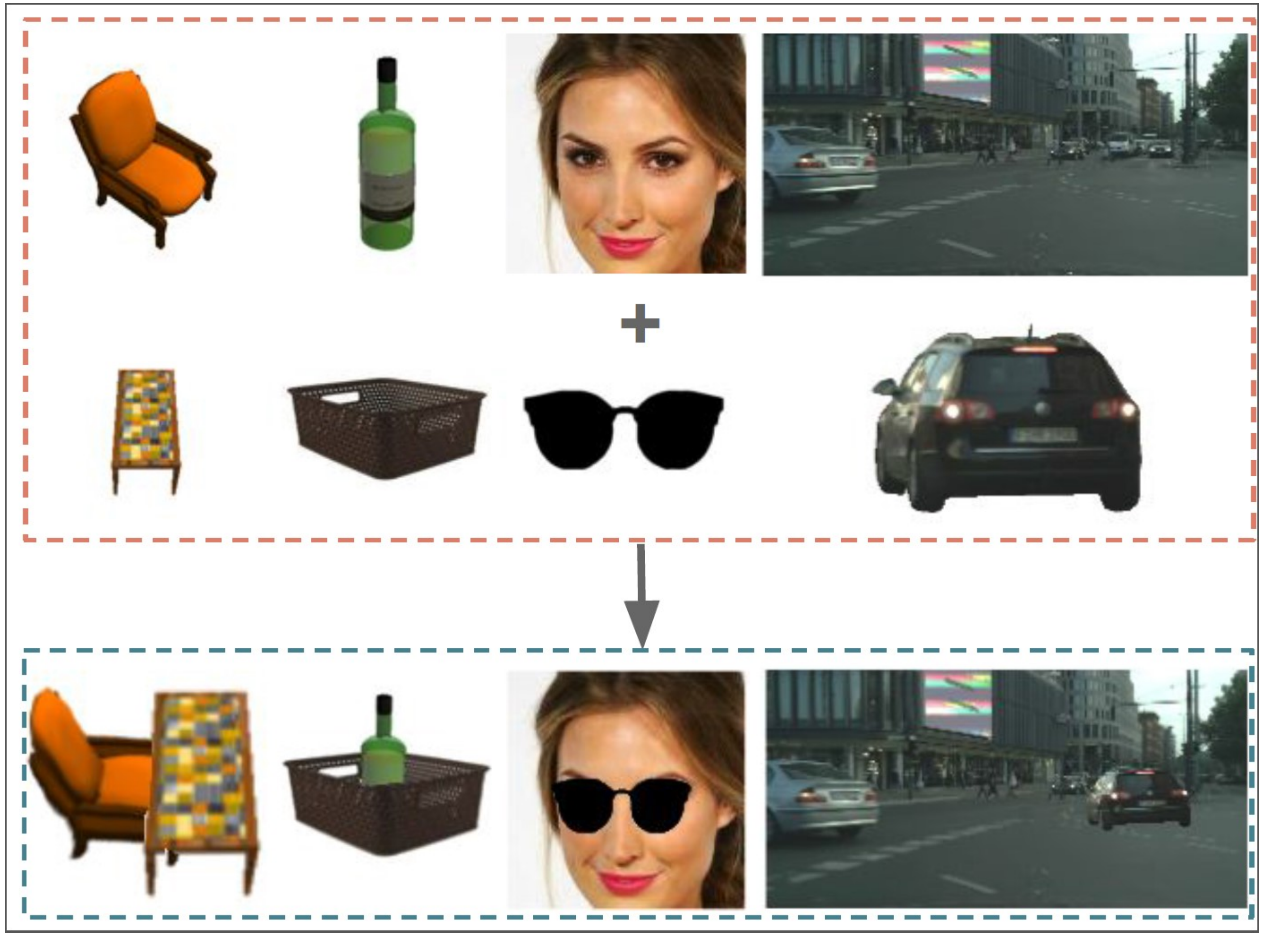


(a)



Motivation

• More general case and more complex interactions



• X + Y => C
• Assumption
The masks for X and Y in C are known

➢ Paired case
Training data contains X, Y and C

➢ Unpaired case
Training data contains only C

Approaches

For paired cases, *GAN not good at transforming objects spatially*

- Spatial Transformer Network (STN)
- Relative Appearance Flow Network (RAFN)

⇒Achieve scale and shift transformation of objects by STN, adjust viewpoint using RAFN

For unpaired cases, cut out X and Y from C

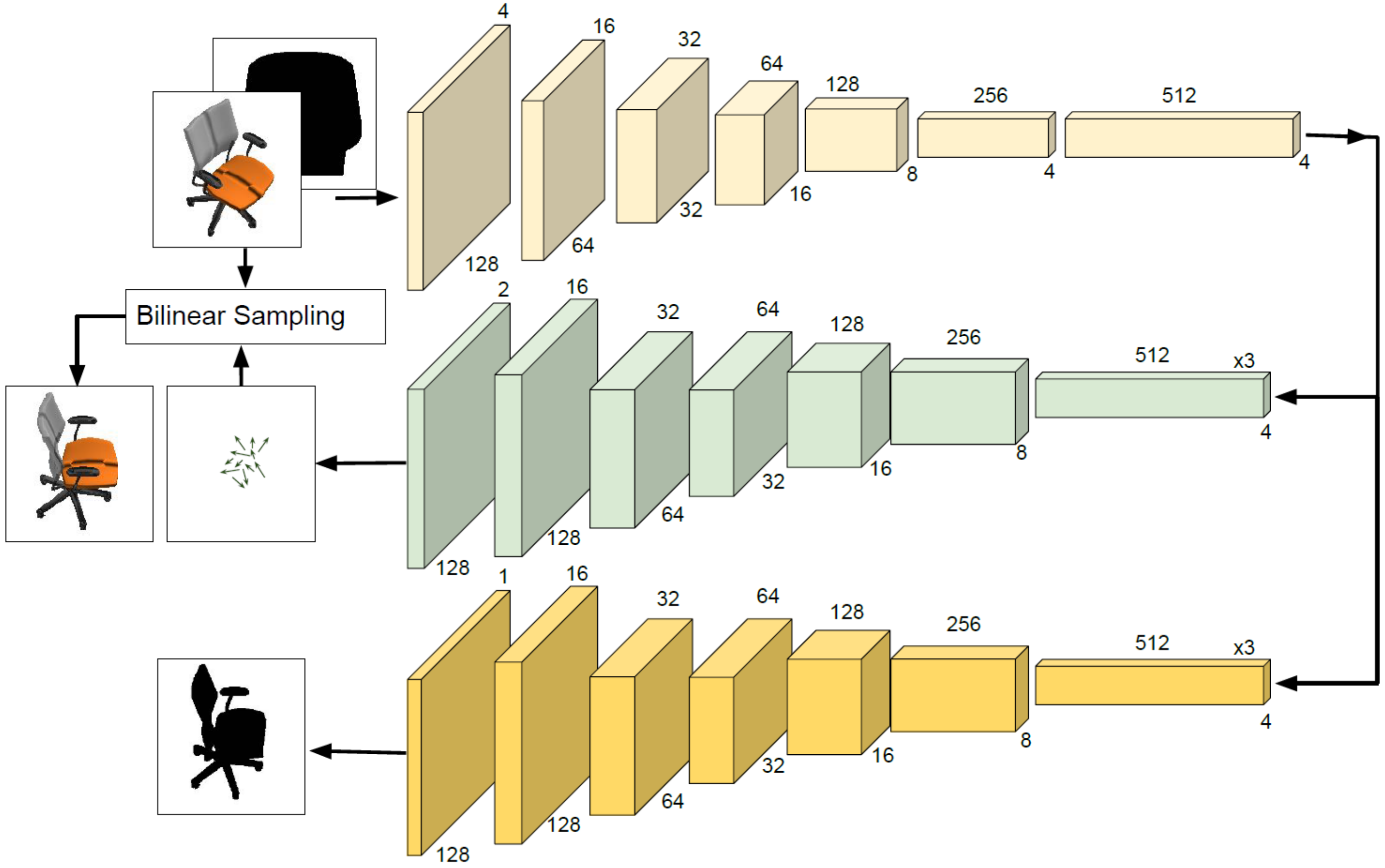
- Self-supervised inpainting network

⇒Obtained well aligned X and Y

Finally, use conditional GAN (CGAN) framework to achieve composition and refinement.

RAFN

- Relative Appearance Flow Network based on encoder-decoder



- Training data:

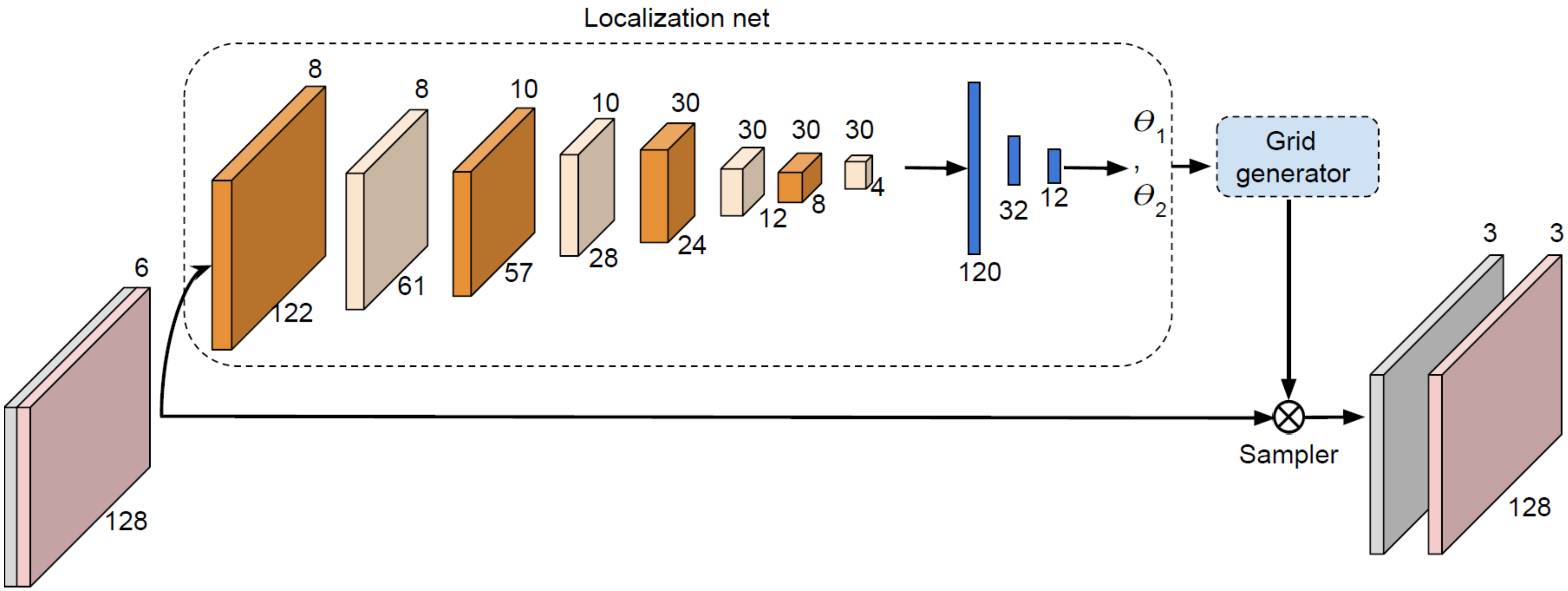
Xr, Y mask, X

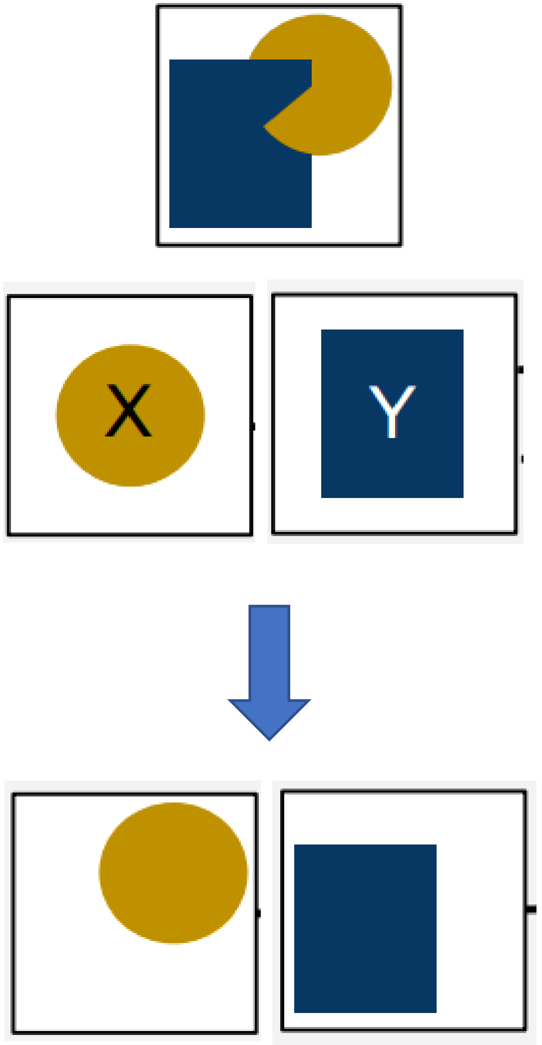
- Loss function

$$\begin{aligned}\mathcal{L}(G_{\text{RAFN}}) &= \mathcal{L}_{L_1}(G_{\text{RAFN}}) + \lambda \mathcal{L}_{\text{BCE}}(G^{M}_{\text{RAFN}}) \\ &= \mathbb{E}_{(x,y)}[\|x - G_{\text{RAFN}}(M^{\text{fg}}_{y}, x^{r})\|_1] \\ &+ \lambda \mathbb{E}_{x}[\hat{M}^{\text{fg}}_{x} \log M^{\text{fg}}_{x} + (1 - \hat{M}^{\text{fg}}_{x}) \log(1 - M^{\text{fg}}_{x})]\end{aligned}$$

STN

- Relative spatial transformer network





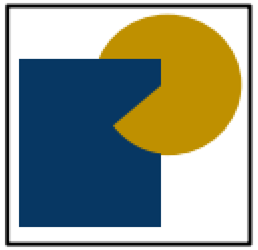
- Training data: X, Y, X^T, Y^T
- Loss function: $\mathcal{L}_{L_1}(\mathrm{STN}) = \mathbb{E}_{(x,y)}[||(x^c, y^c) - (x^T, y^T)||]$

Self-supervised inpainting network (CGAN)

- Training data:

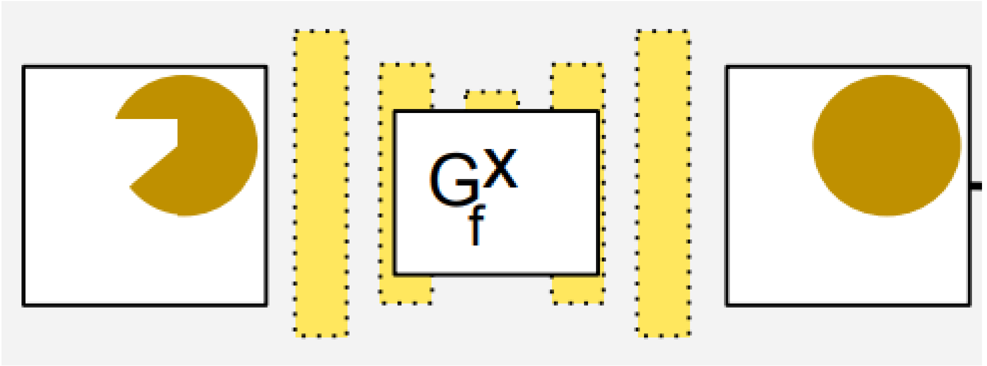
For X: cut out X from C with X mask, zero out pixel values in random area

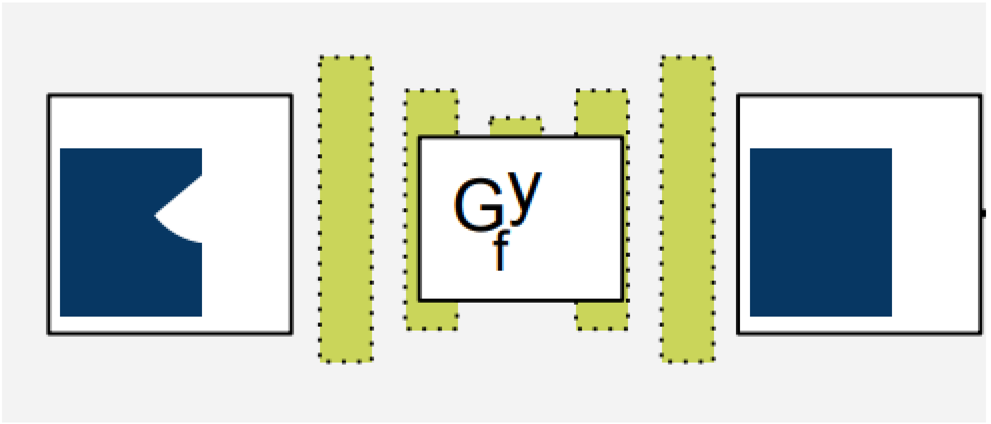
For Y: cut out Y from C with Y mask, zero out the X part



- Loss function

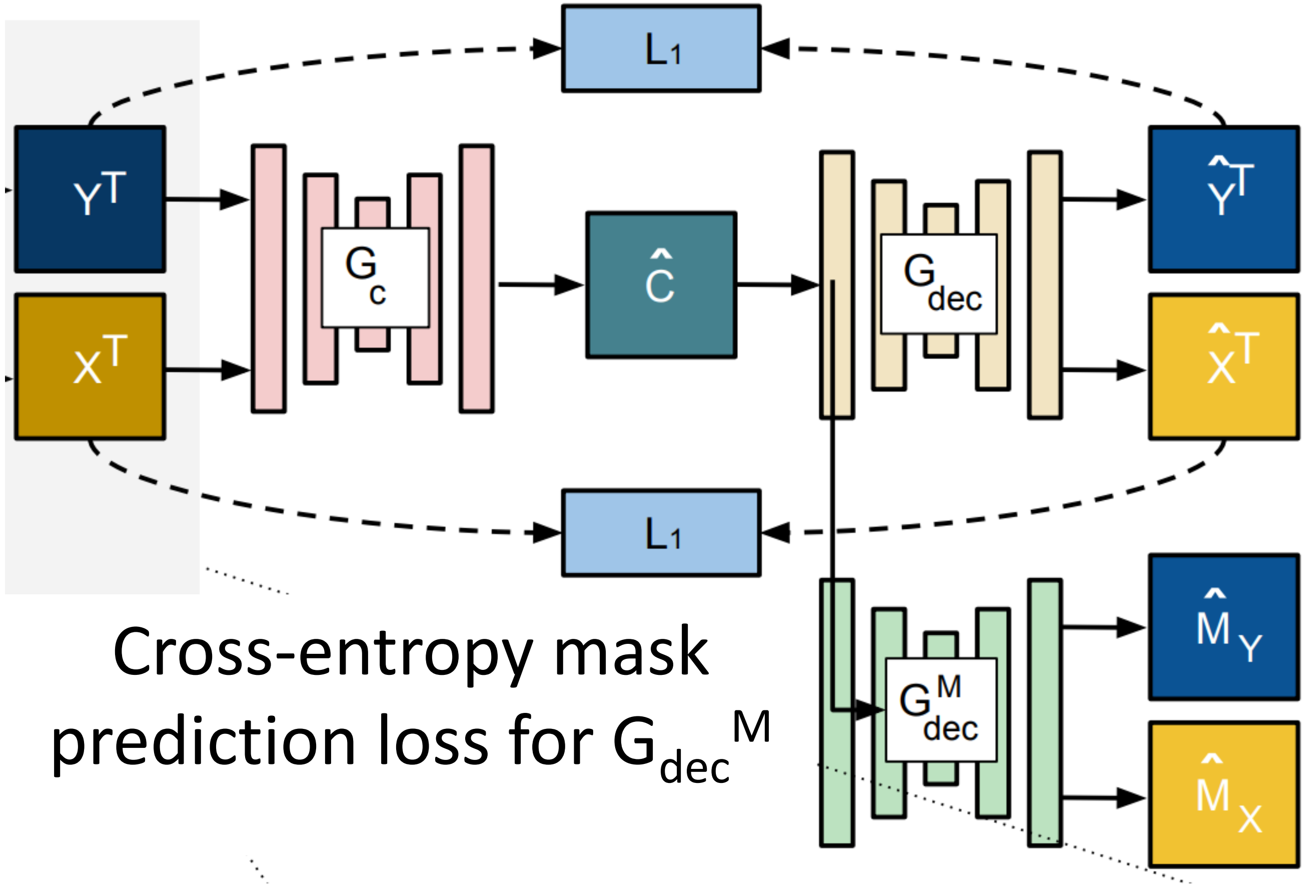
$$\mathcal{L}(G_f) = \mathcal{L}_{L_1}(G_f) + \lambda \mathcal{L}_{cGAN}(G_f, D_f)$$





Supervising composition by decomposition

- Self-consistent Composition-by-Decomposition (CoDe) based on CGAN



Cross-entropy mask prediction loss for G_{dec}^M

- Training data:

X^T, Y^T, Masks, C

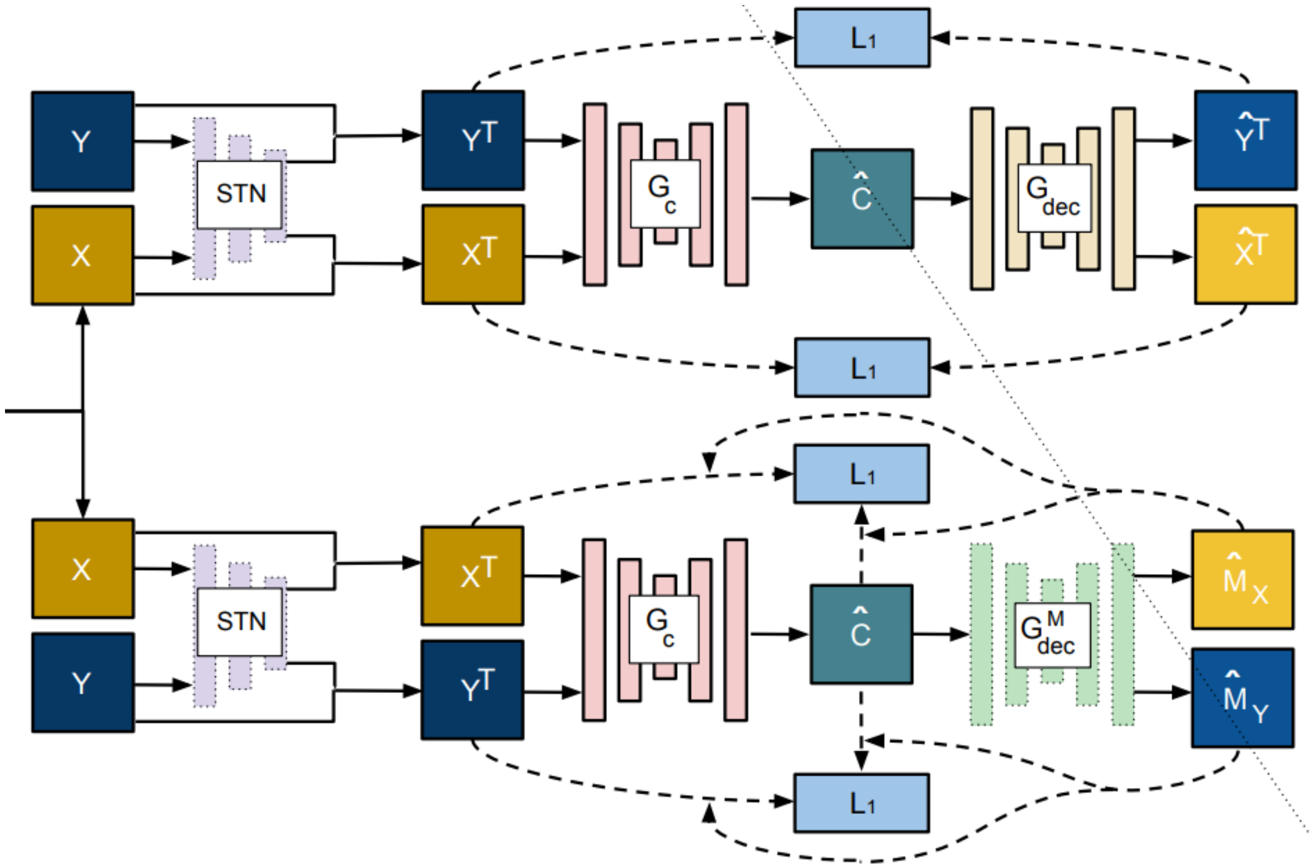
- Loss function

$$
\begin{aligned}
\mathcal{L}_{L_1}(G_c) &= \mathbb{E}_{(x,y,c)}\left[\|c-\hat{c}\|_1\right], \\
\mathcal{L}_{L_1}(G_{dec}) &= \mathbb{E}_{(x,y)}\left[\|(x^T, y^T) - G_{dec}(\hat{c})\|_1\right], \\
\mathcal{L}_{cGAN}(G_c, D_c) &= \mathbb{E}_{(x,y,c)}\left[\log D_c(x^T, y^T, c)\right] \\
&+ \mathbb{E}_{(x,y)}\left[1 - \log D_c(x^T, y^T, \hat{c})\right],
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{L}_{cGAN}(G_{dec}, D_{dec}) = \mathbb{E}_{(x,y)}\big[\log D_{dec}(\hat{c}, x^c) \\
+ \log D_{dec}(\hat{c}, y^c)\big] + \mathbb{E}_{(x,y)}\big[(1 - \log D_{dec}(\hat{c}, \hat{x}^T)) \\
+ (1 - \log D_{dec}(\hat{c}, \hat{y}^T))\big].
\end{aligned}
$$

Example-Specific Meta-Refinement (ESMR)

- Self-consistent Composition-by-Decomposition (CoDe) based on CGAN

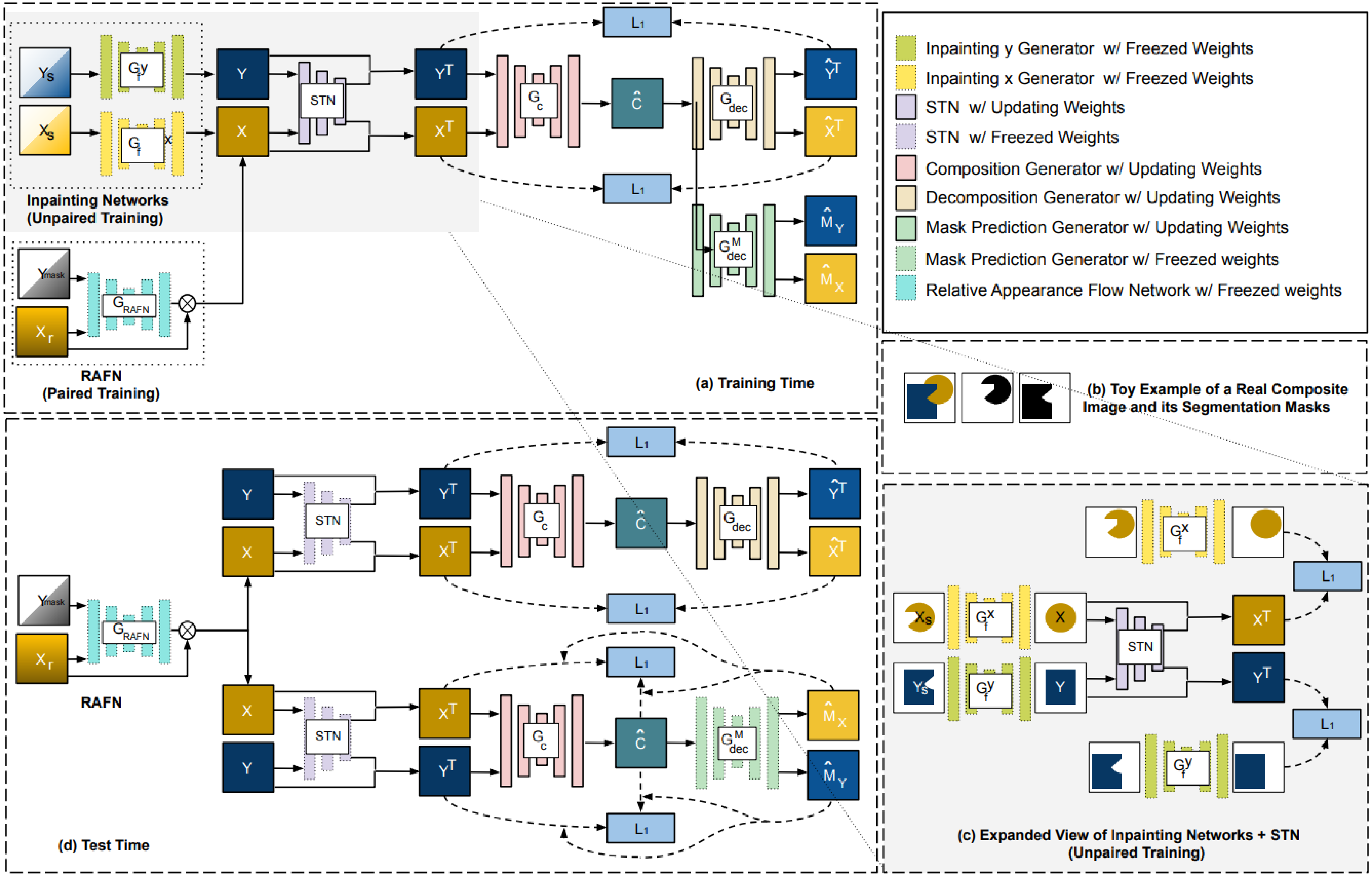


- Loss function

$$
\begin{aligned}
\mathcal{L}(G) \quad = \quad & \lambda(\|\hat{x}^T - x^T\|_1 + \|\hat{M}_x \odot \hat{c} - \hat{M}_x \odot x^T\|_1 \\
+ \quad & \|\hat{y}^T - y^T\|_1 + \|\hat{M}_y \odot \hat{c} - \hat{M}_y \odot y^T\|_1) \\
+ \quad & [\mathcal{L}_{\text{cGAN}}(G_c, D_c) + \mathcal{L}_{\text{cGAN}}(G_{\text{dec}}, D_{\text{dec}})],
\end{aligned}
$$

- Freeze STN, RAFN, and G_{dec}^{M}

- Only refine the weights of CoDe

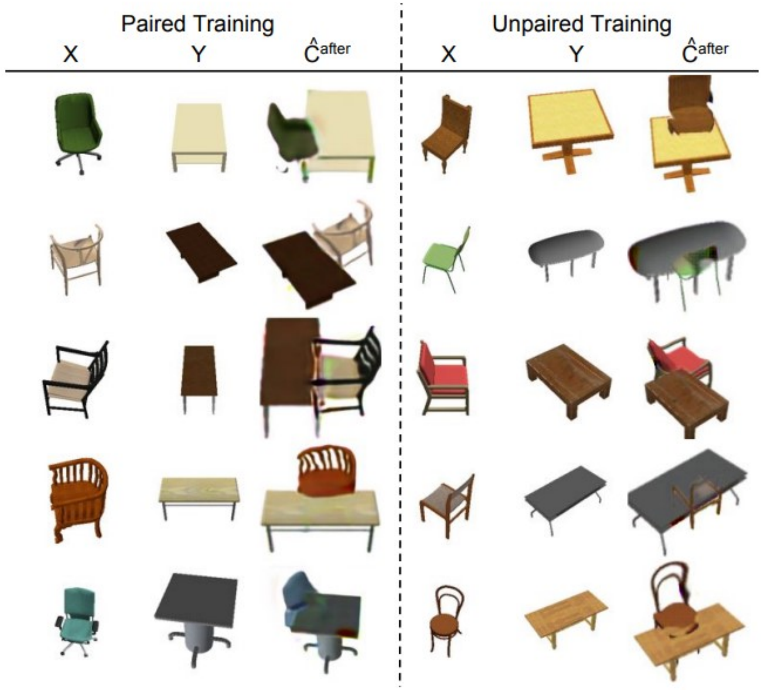


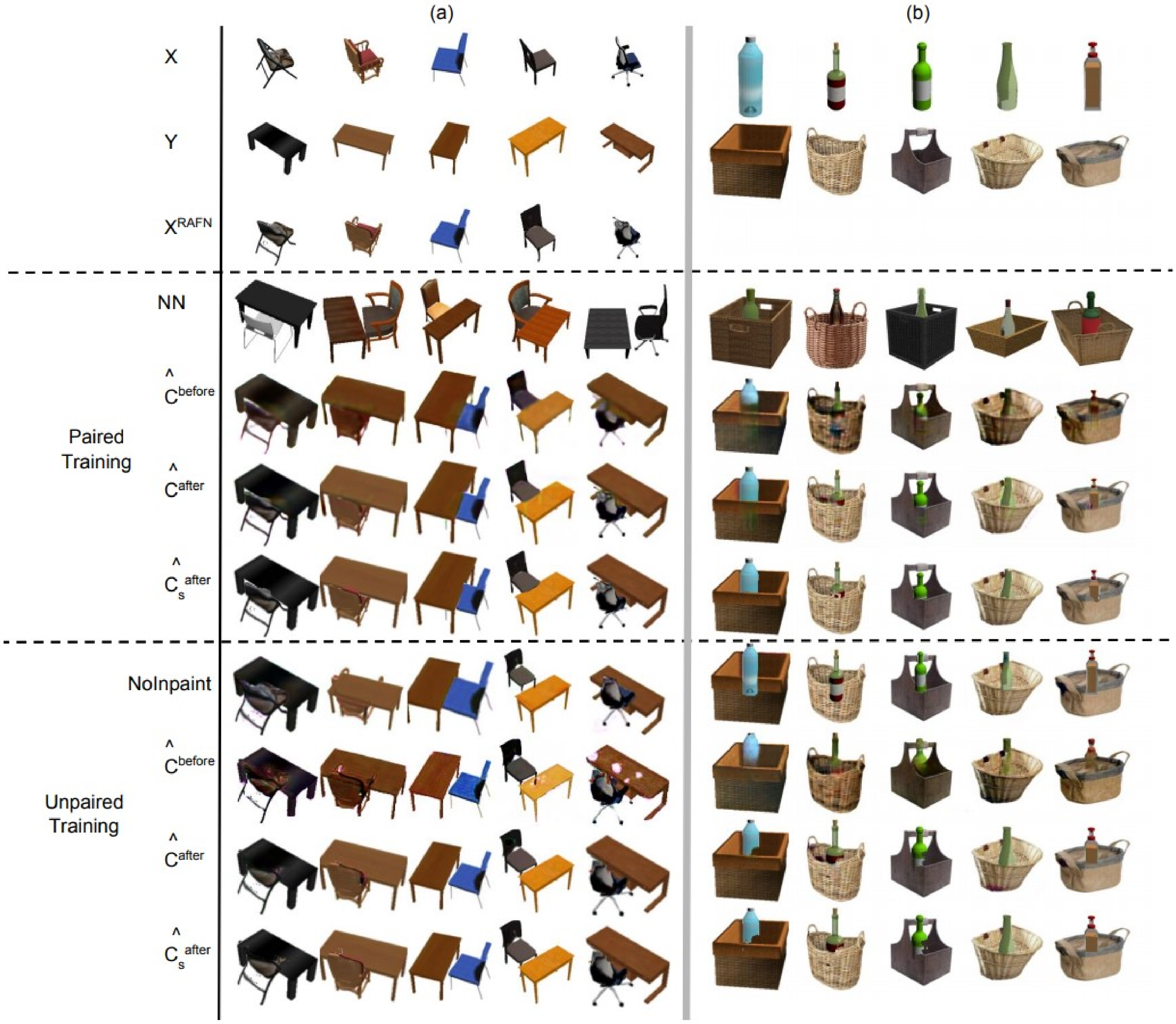
Inpainting y Generator w/ Freezed Weights
Inpainting x Generator w/ Freezed Weights
STN w/ Updating Weights
STN w/ Freezed Weights
Composition Generator w/ Updating Weights
Decomposition Generator w/ Updating Weights
Mask Prediction Generator w/ Updating Weights
Mask Prediction Generator w/ Freezed weights
Relative Appearance Flow Network w/ Freezed weights
Inpainting Networks (Unpaired Training)
RAFN (Paired Training)
(a) Training Time
(b) Toy Example of a Real Composite Image and its Segmentation Masks
(c) Expanded View of Inpainting Networks + STN (Unpaired Training)
(d) Test Time
RAFN
STN
L1

Results:

Table 1. AMT user evaluation comparing components of our model on the synthetic datasets. 2nd column: number of test images, 3rd column: % preferences to after vs. before refinement, 4th column: % preferences to paired training vs. unpaired.

Inputs	# test images	after-vs-before refinement	paired-vs-unpaired
Chair-Table	90	71.3%	57%
Basket-Bottle	45	64.2%	57%





Results:

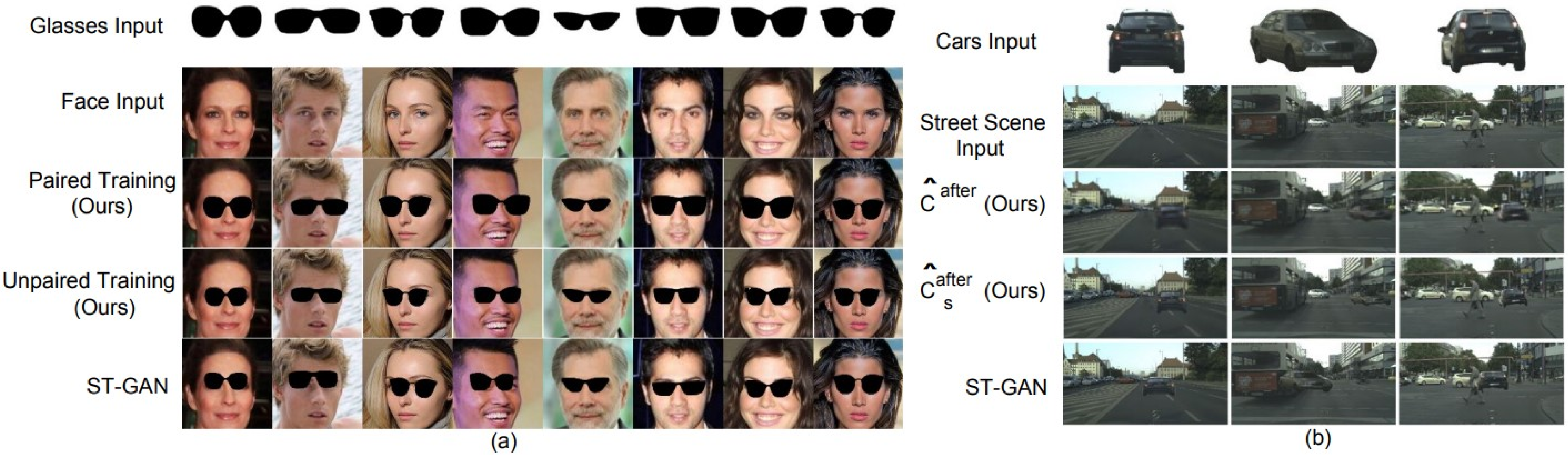
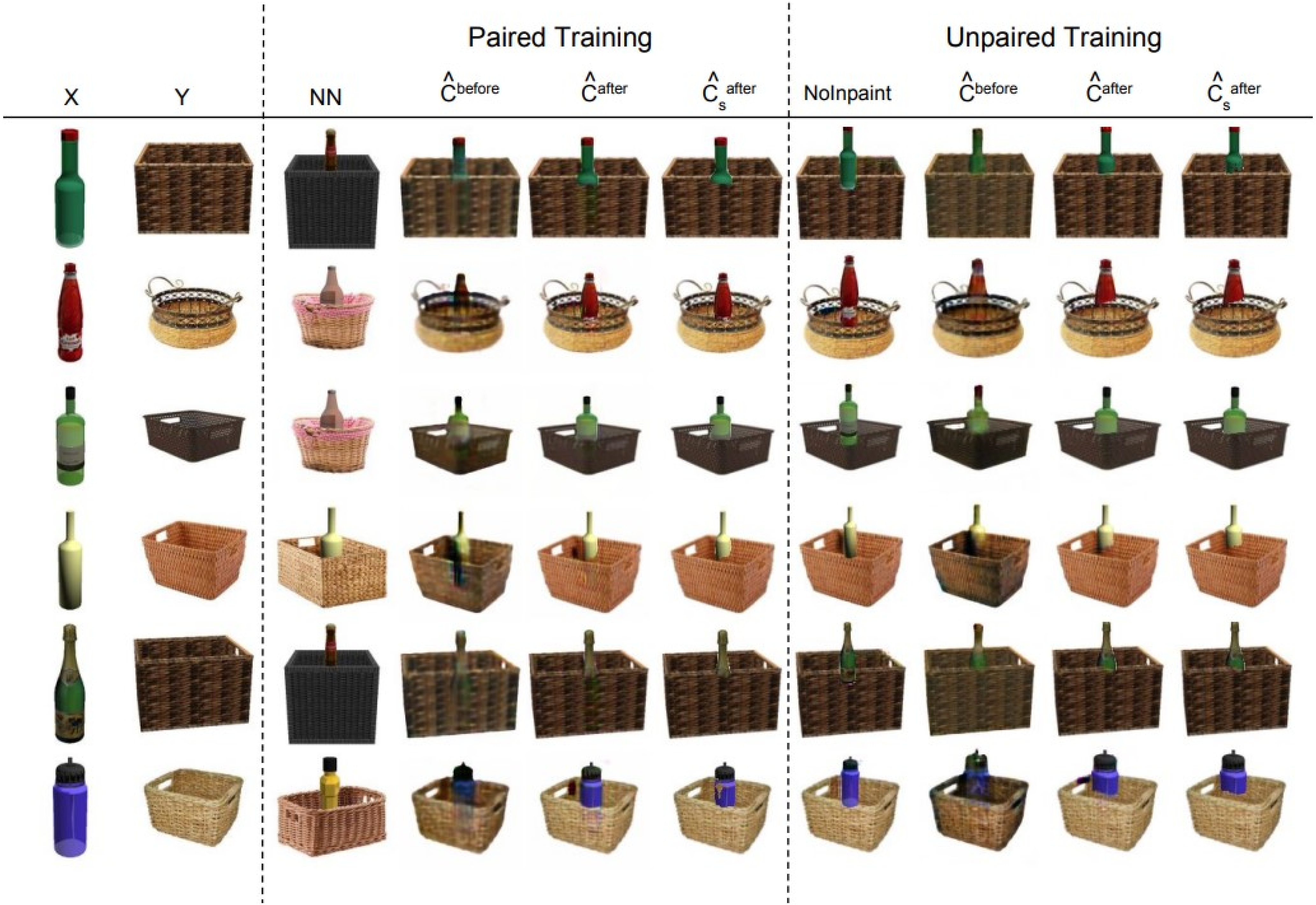
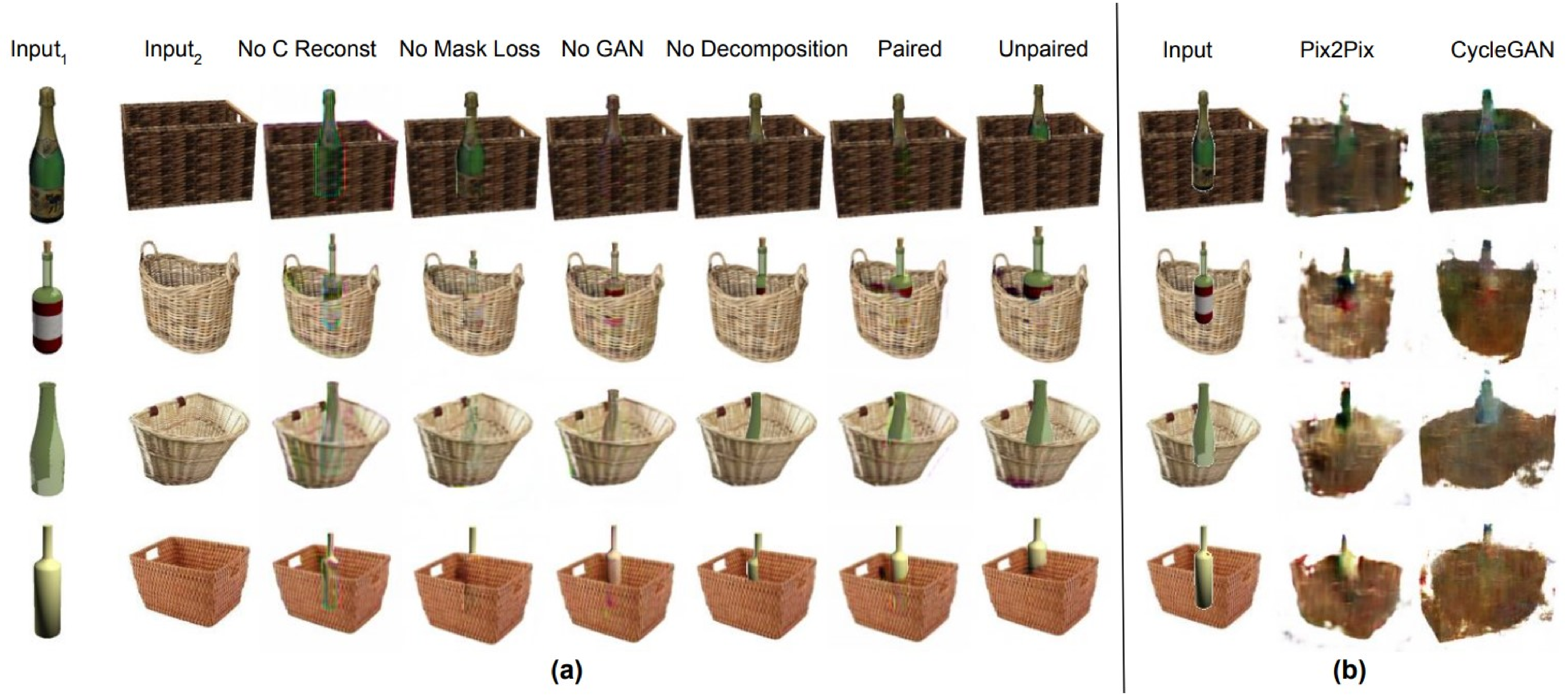


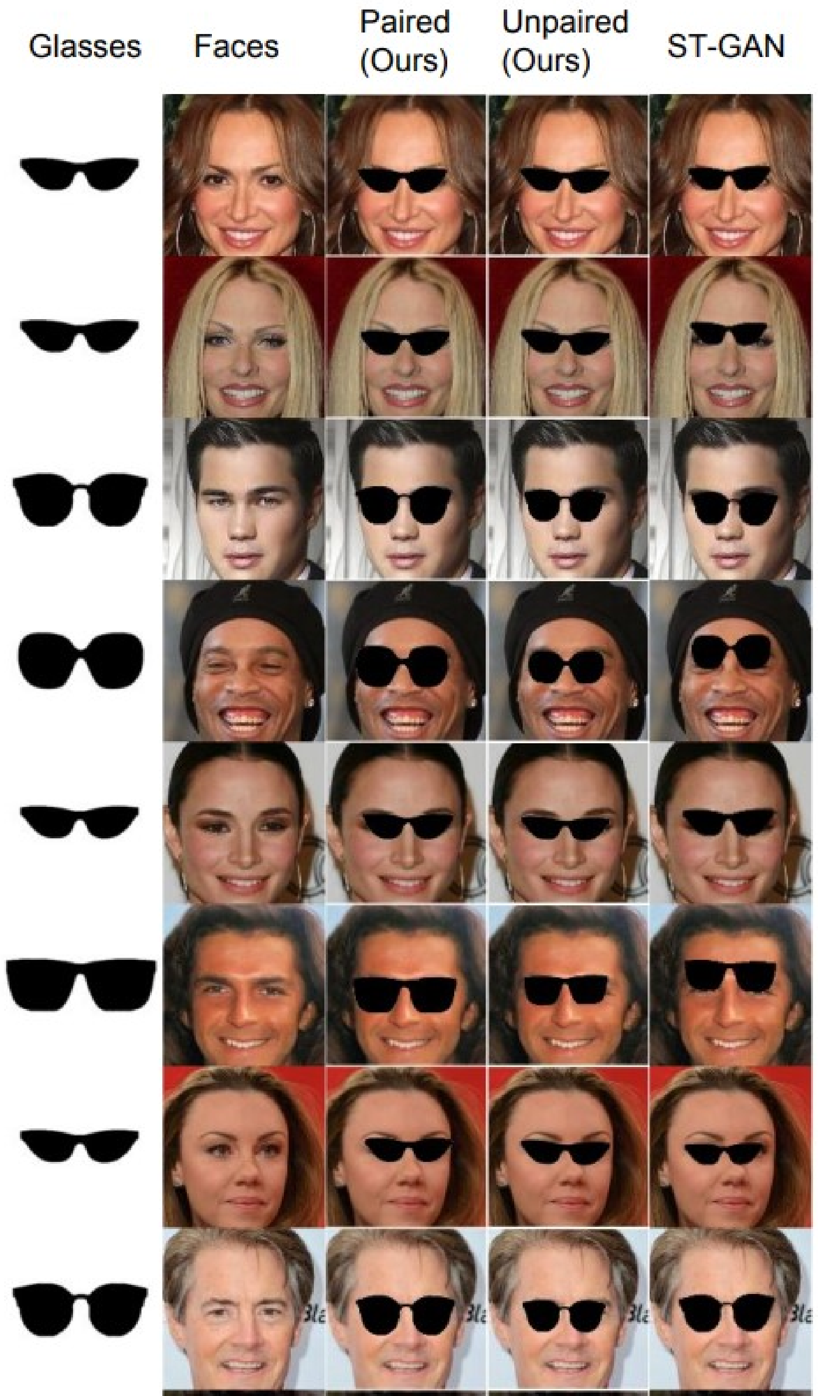
Figure 5. (a) Test examples for the face-sunglasses composition task. *Top two rows*: input sunglasses and face images, *3rd and 4th rows*: the output of our compositional GAN for the paired and unpaired models, respectively, *Last row*: images generated by the ST-GAN [16] model, (b) Test examples for the street scene-car composition task. *Top two rows*: input cars and street scenes, *3rd and 4th rows*: the output of our compositional GAN after the meta-refinement approach. Here, $\hat{c}^{\text{after}}$ shows the output of the composition generator and $\hat{c}_{\text{s}}^{\text{after}}$ represents the summation of the masked transposed inputs, *Last row*: images generated by ST-GAN.



Paired Training
Unpaired Training
X
Y
NN
$\hat{C}^{before}$
$\hat{C}^{after}$
$\hat{C}_s^{after}$
NoInpaint
$\hat{C}^{before}$
$\hat{C}^{after}$
$\hat{C}_s^{after}$


Input1
Input2
No C Reconst
No Mask Loss
No GAN
No Decomposition
Paired
Unpaired
Input
Pix2Pix
CycleGAN
(a)
(b)


Cars
Street Scenes
$\hat{C}^{before}$
$\hat{C}^{after}$
$\hat{C}_s^{after}$
ST-GAN
NN


Glasses
Faces
Paired (Ours)
Unpaired (Ours)
ST-GAN

Thanks for your attention