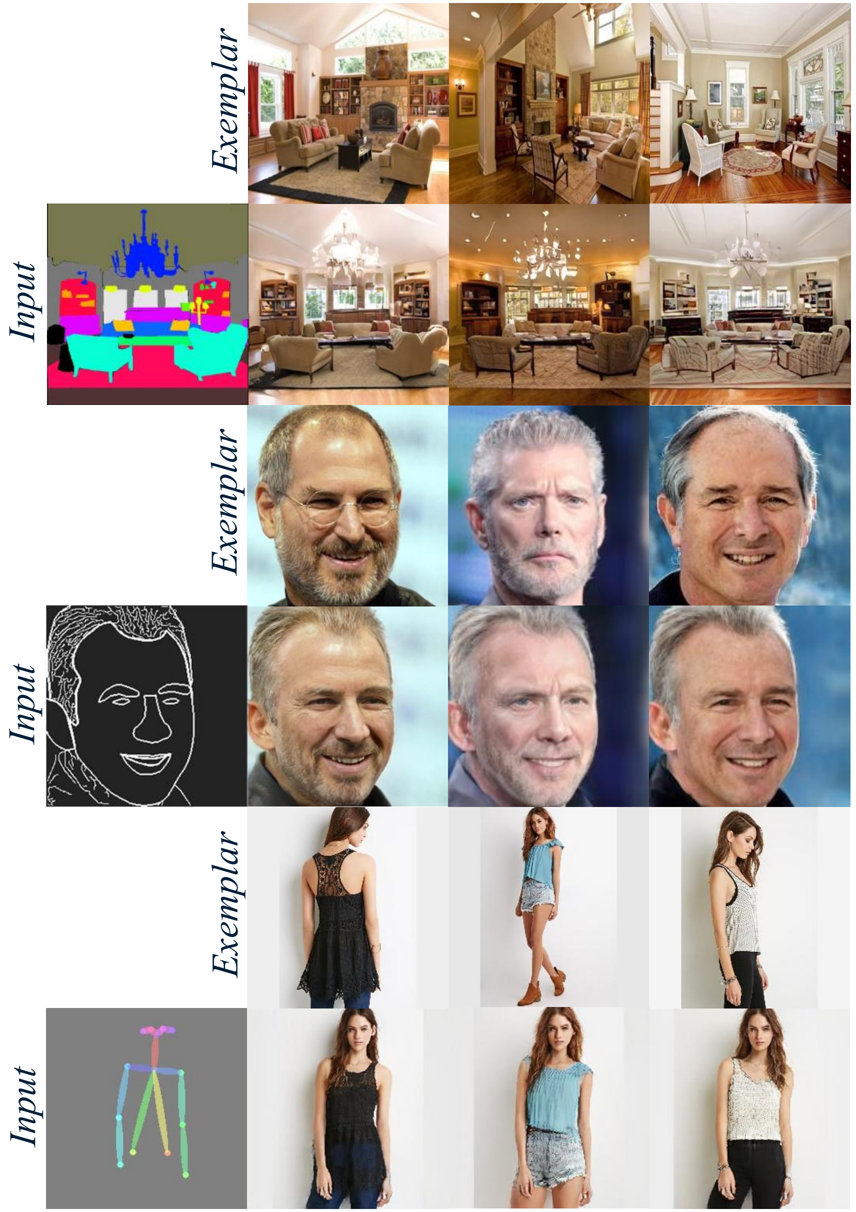
Cross-domain Correspondence Learning for Exemplar-based Image Translation (CVPR 2020)

Microsoft Research Asia

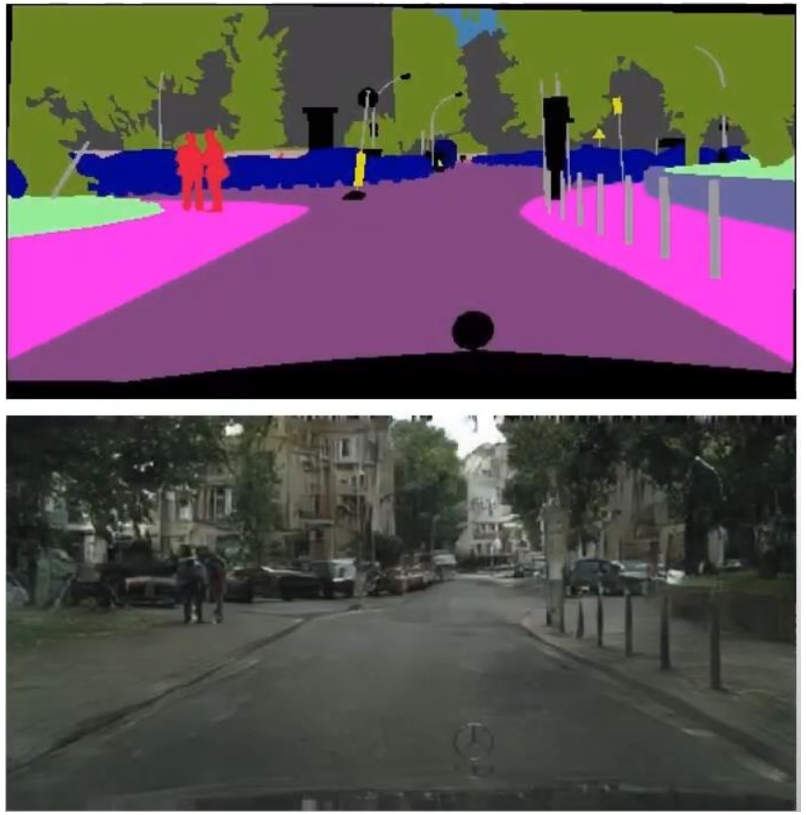


Section

Background

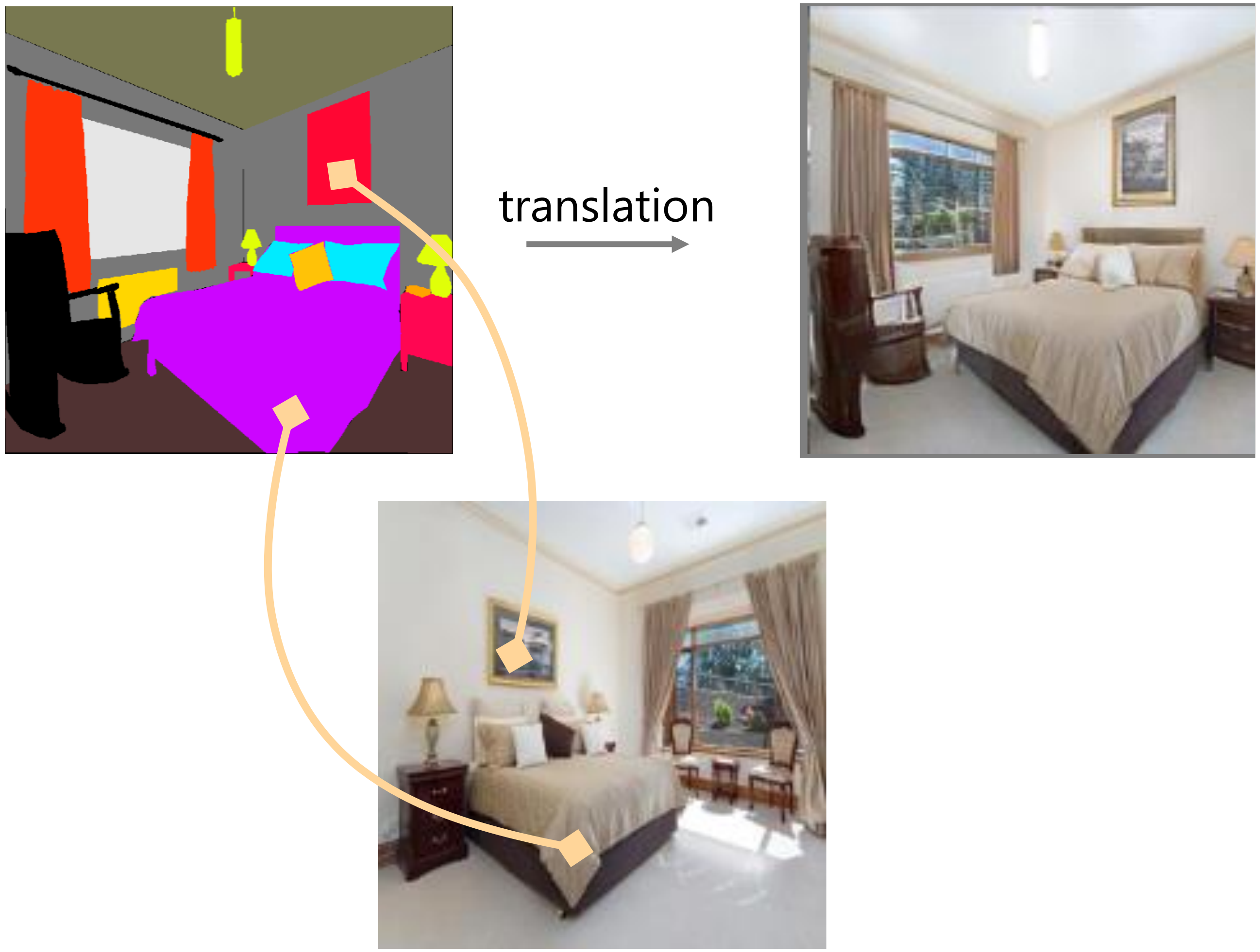
Motivation

(a) edges ↔ shoes



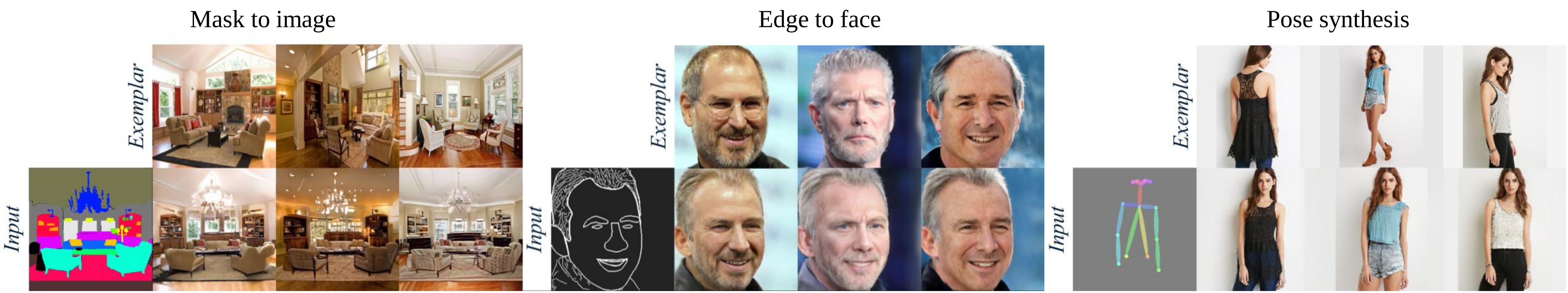
- Lack of controllability
- Compromised image quality for complex scenes

A more user-friendly approach

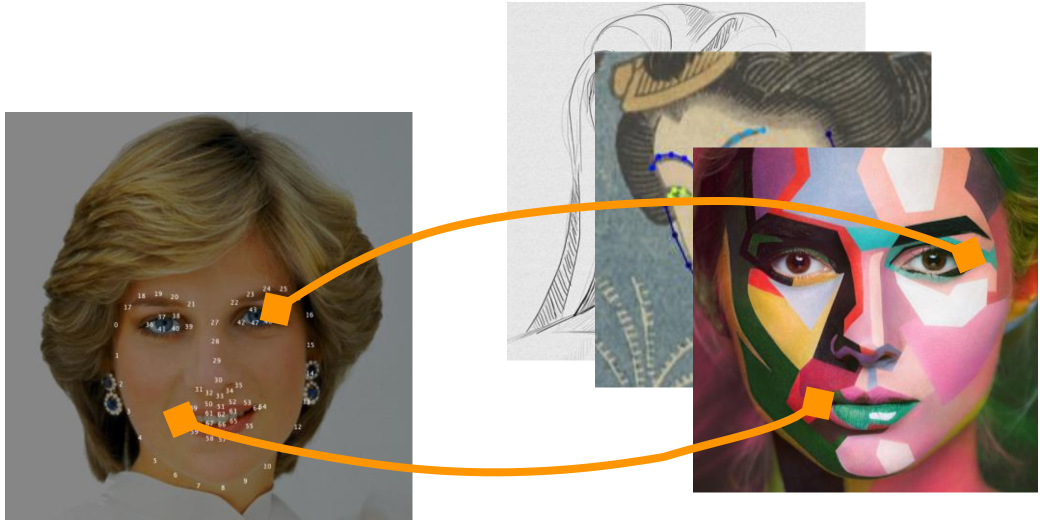


General solution

- Our method transfers the fine structures from the exemplar
- Applicable to various tasks



Challenge: weakly supervised learning



How to establish correspondence for heterogeneous images?

What is the desired translation output given an exemplar?

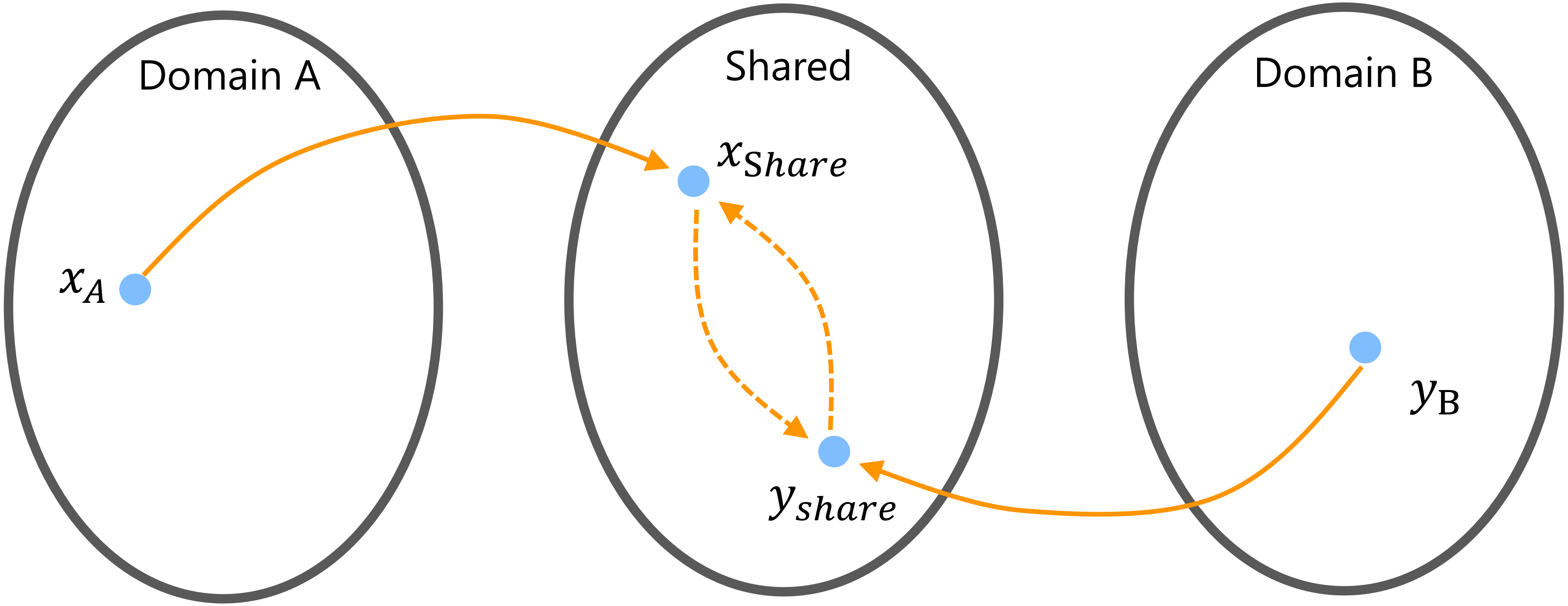
Facilitate each other

Section

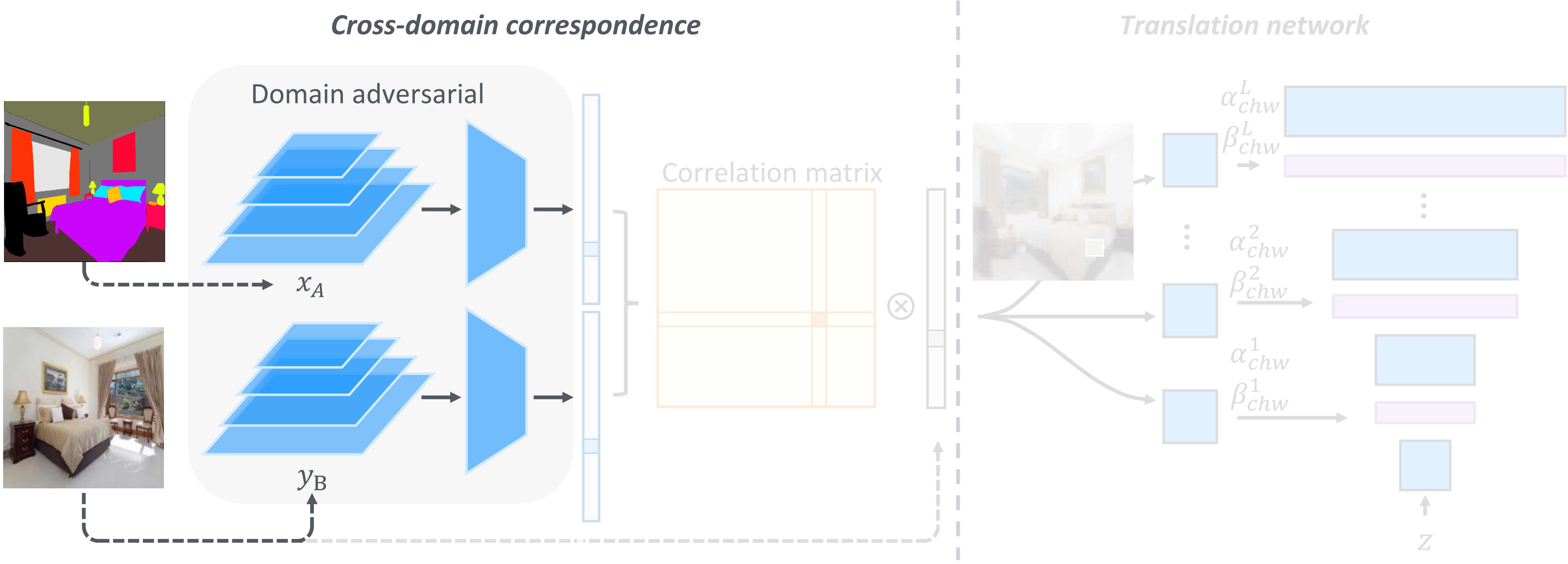
Method

basic idea

- find an intermediate domain which is suitable for dense correspondence

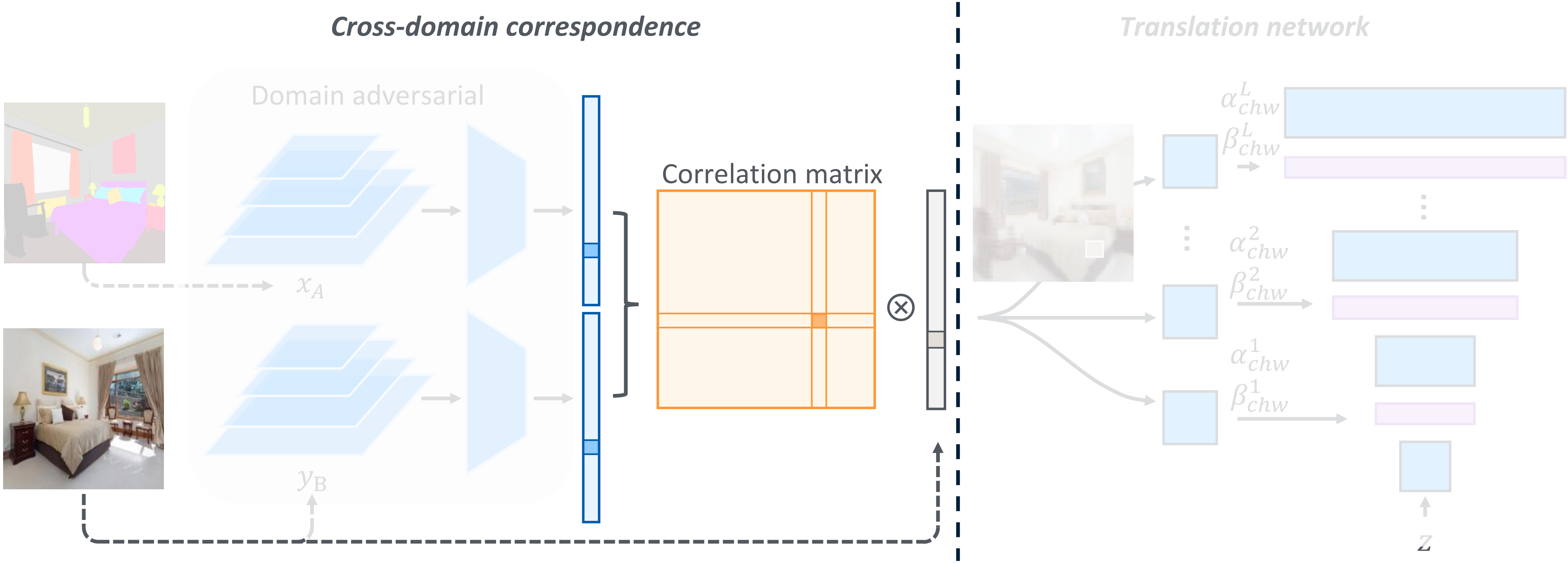


Framework



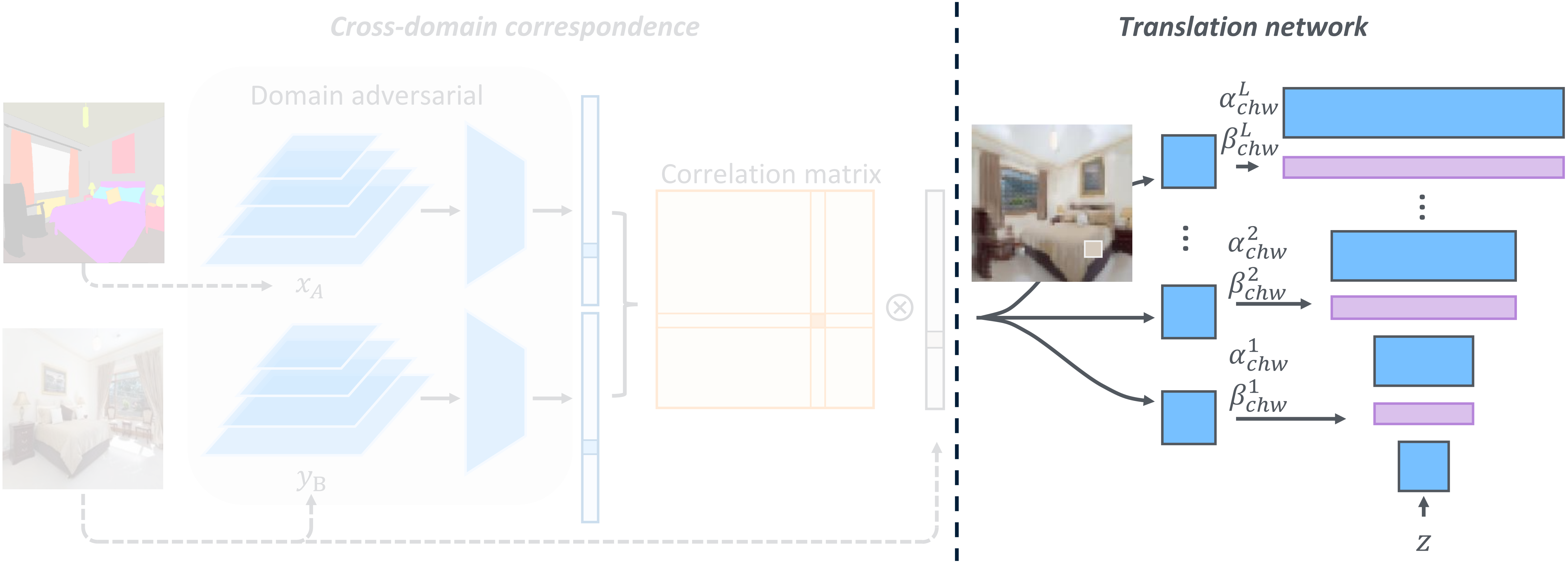
Cross-domain correspondence
Translation network
Domain adversarial
Correlation matrix
x_A
y_B
α^L_{chw}
β^L_{chw}
α^2_{chw}
β^2_{chw}
α^1_{chw}
β^1_{chw}
z

Framework



Cross-domain correspondence
Translation network
Domain adversarial
Correlation matrix
x_A
y_B
α^L_{chw}
β^L_{chw}
α^2_{chw}
β^2_{chw}
α^1_{chw}
β^1_{chw}
z

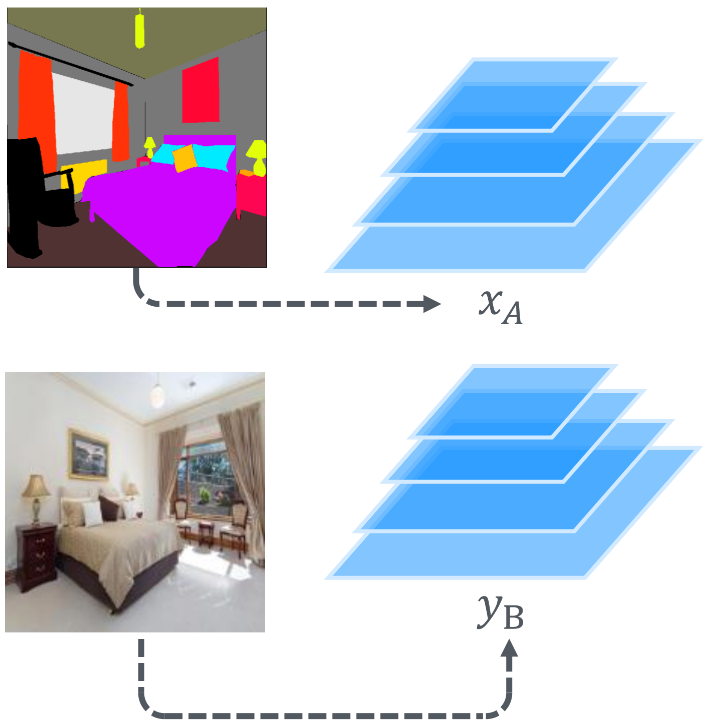
Framework



Cross-domain correspondence
Domain adversarial
x_A
y_B
Correlation matrix
Translation network
α^L_{chw}
β^L_{chw}
α^2_{chw}
β^2_{chw}
α^1_{chw}
β^1_{chw}
z

Cross-domain correspondence network

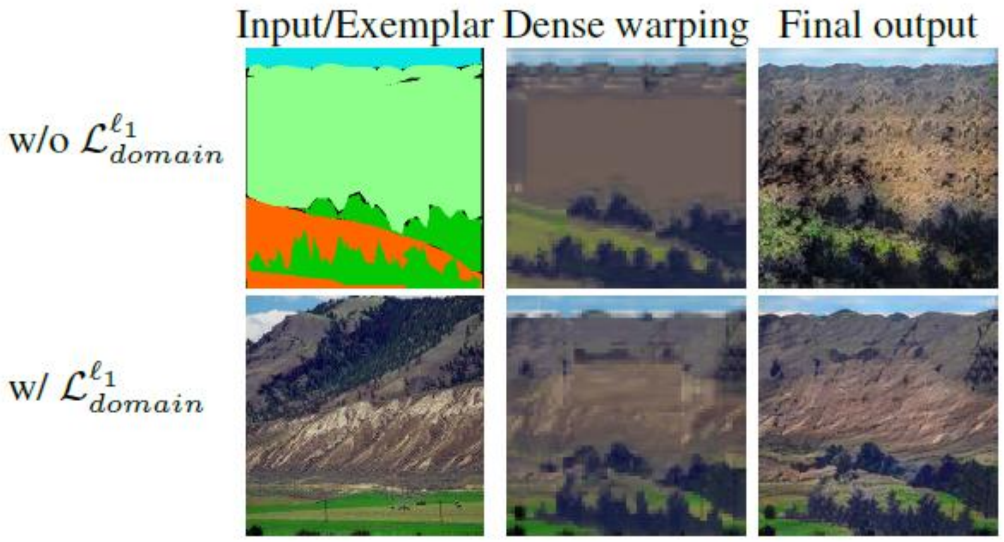
Domain adversarial



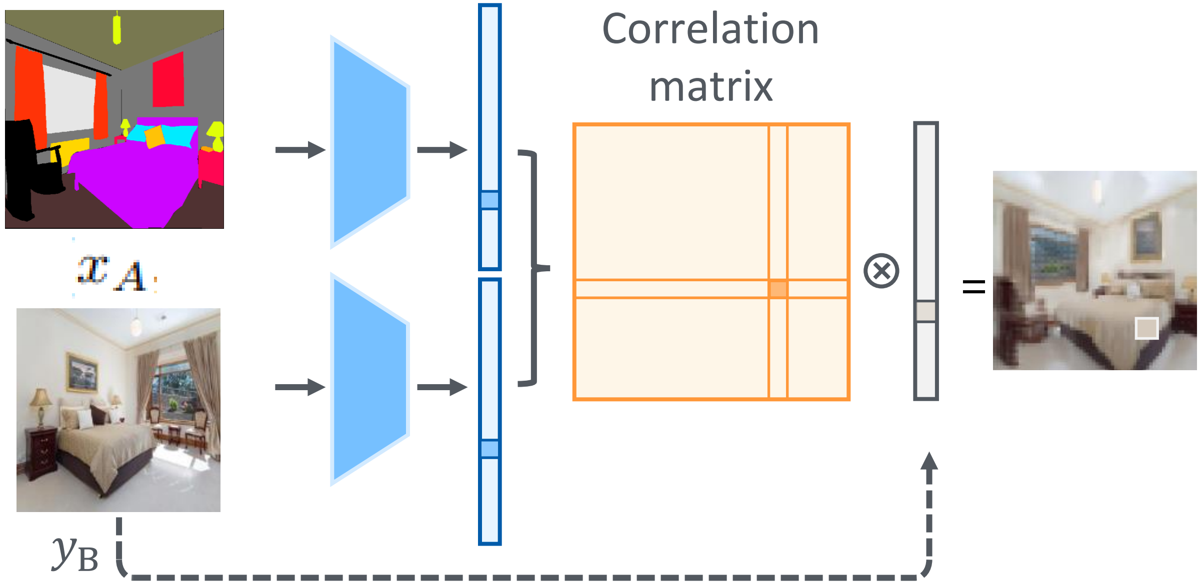
$$x_S = \mathcal{F}_{A\to S}(x_A; \theta_{\mathcal{F},A\to S}),$$
$$y_S = \mathcal{F}_{B\to S}(y_B; \theta_{\mathcal{F},B\to S}).$$

Domain alignment loss: the feature embedding should lie in the same domain

$$\mathcal{L}_{domain}^{\ell_1} = \|\mathcal{F}_{A\to S}(x_A) - \mathcal{F}_{B\to S}(x_B)\|_1$$



Cross-domain correspondence network



Within domain correspondence:

$$\mathcal{M}(u,v) = \frac{\hat{x}_S(u)^T \hat{y}_S(v)}{\|\hat{x}_S(u)\| \; \|\hat{y}_S(v)\|},$$

Soft exemplar warping:

$$r_{y\to x}(u) = \sum_v \underset{v}{\mathrm{softmax}}(\alpha \mathcal{M}(u,v)) \cdot y_B(v).$$

Cyclic exemplar warping

$$r_{y\to x\to y}(v) = \textstyle\sum_u \mathrm{softmax}_u(\alpha \mathcal{M}(u,v)) \cdot r_{y\to x}(u)$$

Channel wise centralized feature
X^s(u) = Xs(u) – mean(Xs(u))
Y^s(v) = Ys(v) – mean(Ys(v))

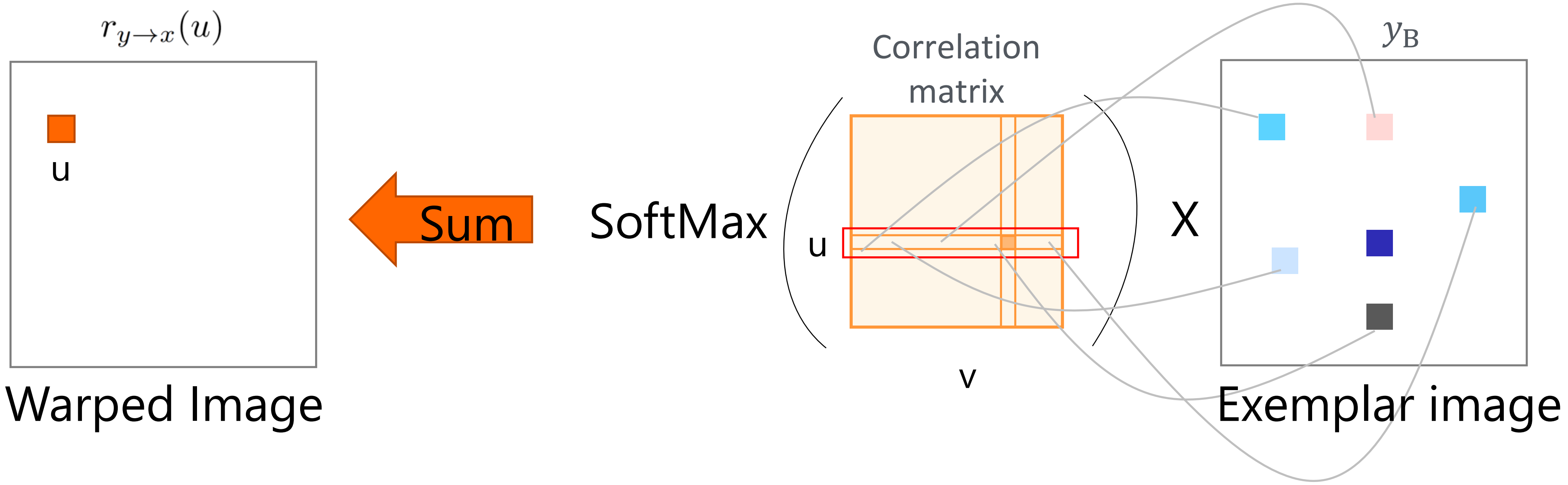
Cross-domain correspondence network

Soft exemplar warping:

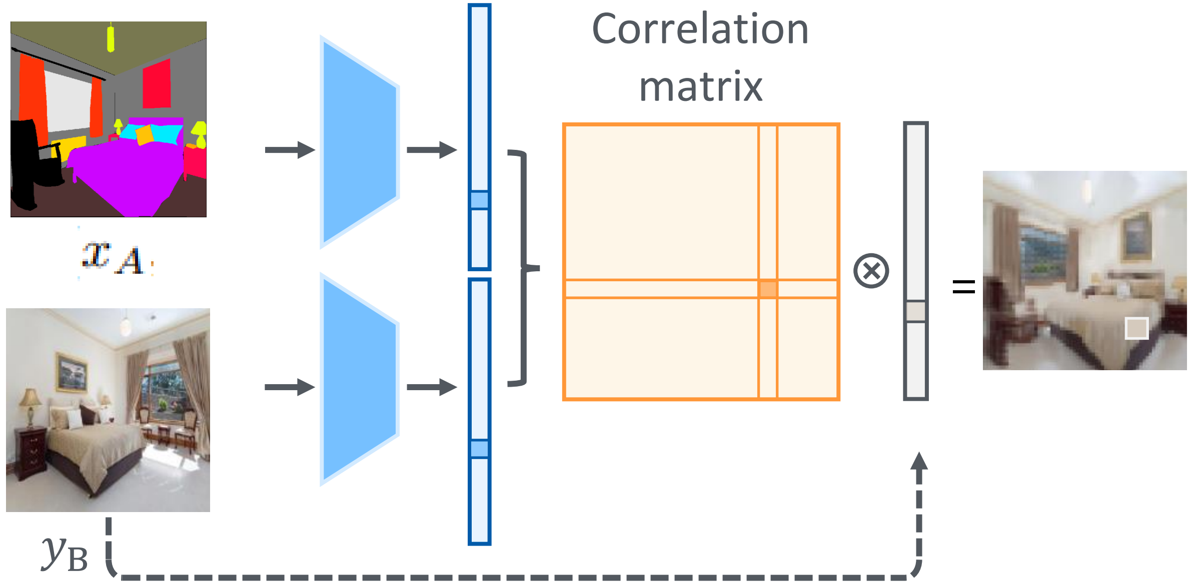
$$r_{y\to x}(u) = \sum_v \underset{v}{\mathrm{softmax}}(\alpha \mathcal{M}(u,v)) \cdot y_B(v).$$

Cyclic exemplar warping

$$r_{y\to x\to y}(v) = \textstyle\sum_u \mathrm{softmax}_u(\alpha \mathcal{M}(u,v)) \cdot r_{y\to x}(u)$$



Cross-domain correspondence network

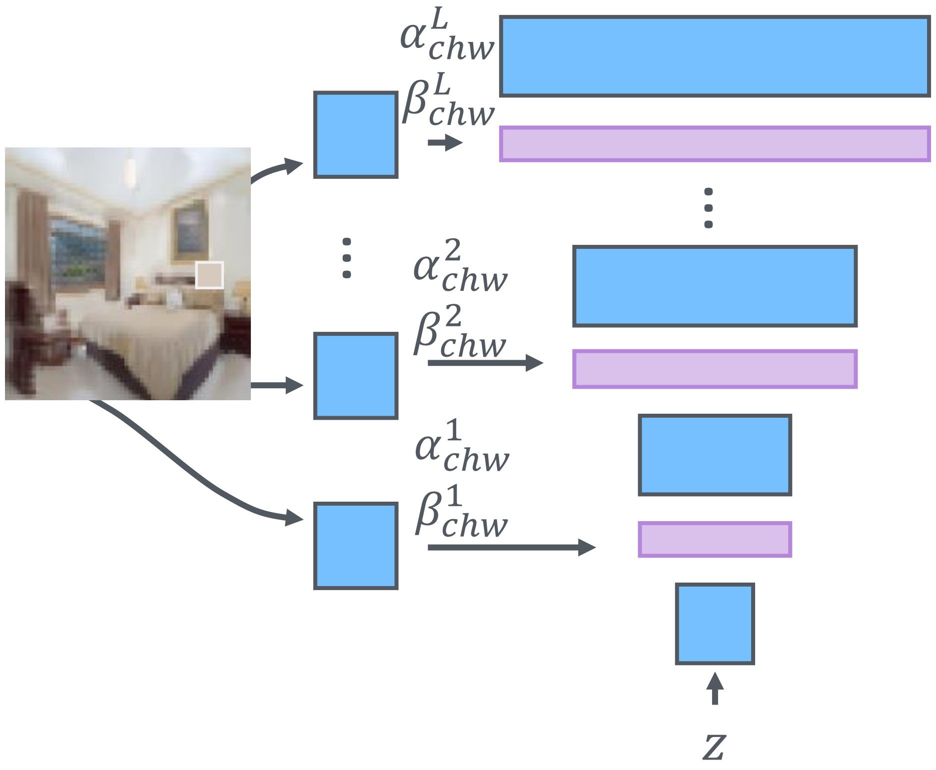


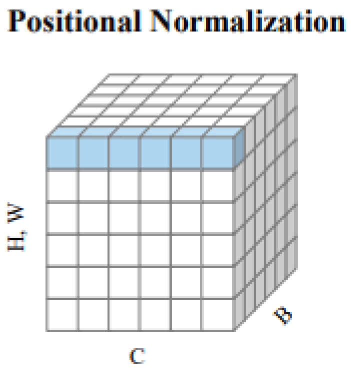
Correspondence regularization:

$$\mathcal{L}_{reg} = \left\| r_{y \to x \to y} - y_B \right\|_1 ,$$

Translation network

Translation network





Spatial variant style injection:

$$\alpha^i_{h,w}(r_{y\to x}) \times \frac{F^i_{c,h,w} - \mu^i_{h,w}}{\sigma^i_{h,w}} + \beta^i_{h,w}(r_{y\to x}),$$

Positional normalization

Style encoder:

$$\alpha^i, \beta^i = \mathcal{T}_i(r_{y\to x}; \theta_\mathcal{T}).$$

Image Translation Output:

$$\hat{x}_B = \mathcal{G}(z, \mathcal{T}_i(r_{y\to x}; \theta_\mathcal{T}); \theta_\mathcal{G}).$$

Translation network

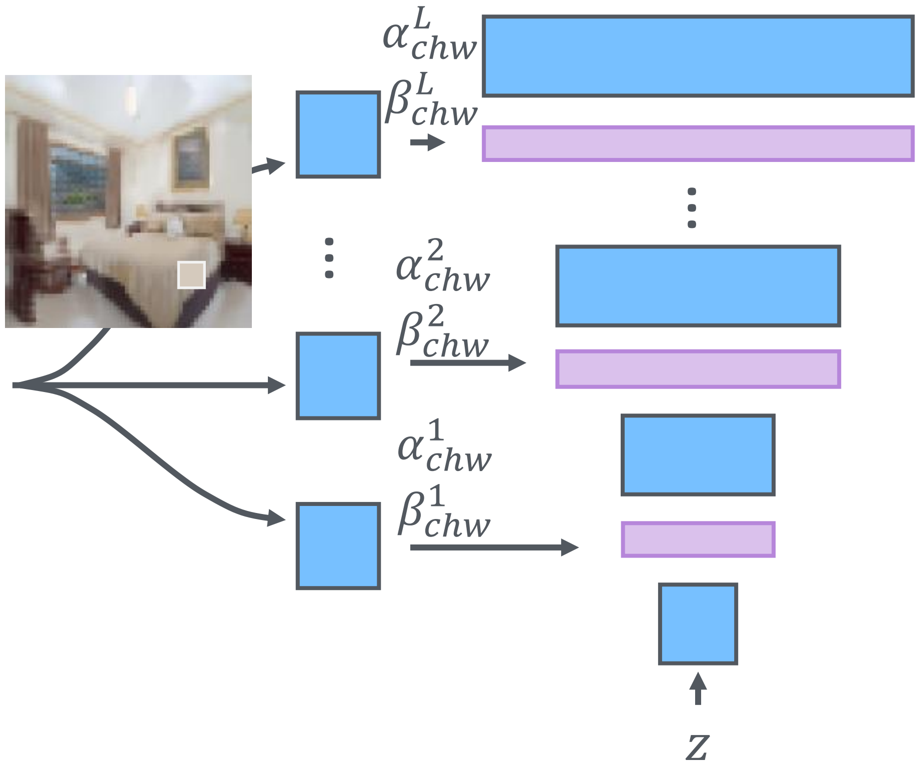
Pseudo exemplar loss:

$$\mathcal{L}_{feat} = \sum_l \lambda_l \left\| \phi_l(\mathcal{G}(x_A, x'_B)) - \phi_l(x_B) \right\|_1 ,$$

Pseudo exemplar pairs

Translation network

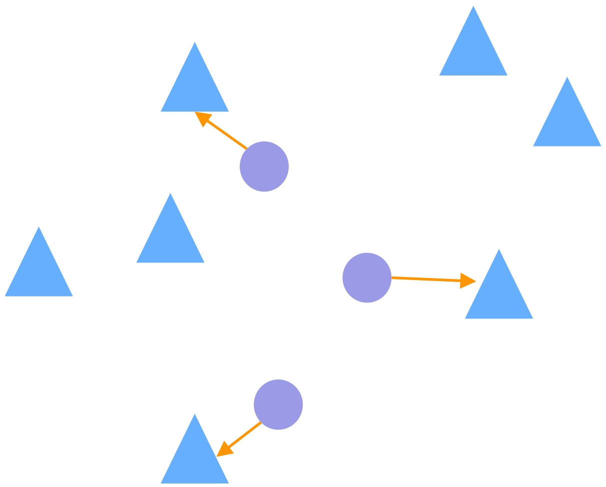
Translation network



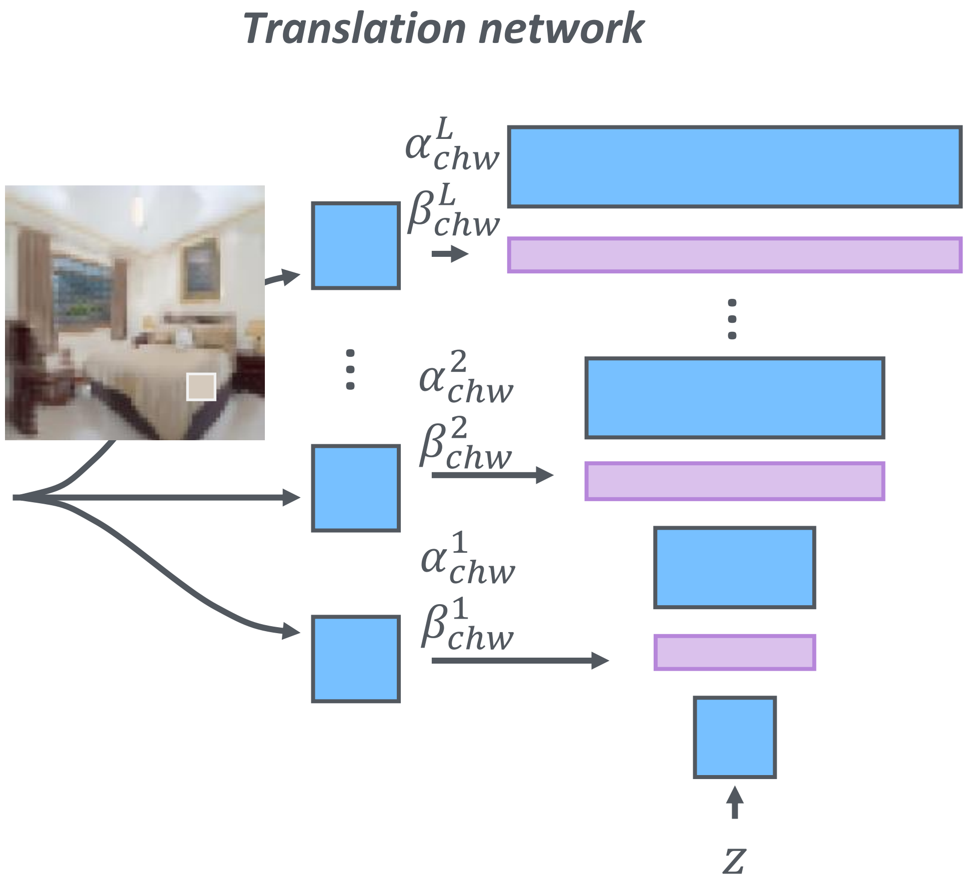
Contextual loss: let the output to mimic the appearance of the semantically corresponding patches from the exemplar.

$$\mathcal{L}_{context} = \sum_l \omega_l \left[-\log \left(\frac{1}{n_l} \sum_i \max_j A^l(\phi_i^l(\hat{x}_B), \phi_j^l(y_B)) \right) \right],$$

pairwise affinities between the features of output and their closest features of the exemplar



Translation network



Perceptual loss: the output should maintain the semantics as the input

$$\mathcal{L}_{perc} = \left\| \phi_l(\hat{x}_B) - \phi_l(x_B) \right\|_1 .$$

Adversarial loss: make the output as realistic as possible

$$\mathcal{L}^{\mathcal{D}}_{adv} = -\mathbb{E}[h(\mathcal{D}(y_B))] - \mathbb{E}[h(-\mathcal{D}(\mathcal{G}(x_A, y_B)))]$$
$$\mathcal{L}^{\mathcal{G}}_{adv} = -\mathbb{E}[\mathcal{D}(\mathcal{G}(x_A, y_B))],$$

Total loss

$$
\mathcal{L}_\theta = \min_{\mathcal{F},\mathcal{T},\mathcal{G}} \max_{\mathcal{D}} \psi_1 \mathcal{L}_{feat} + \psi_2 \mathcal{L}_{perc} + \psi_3 \mathcal{L}_{context} + \psi_4 \mathcal{L}_{adv}^{\mathcal{G}} + \psi_5 \mathcal{L}_{domain}^{\ell_1} + \psi_6 \mathcal{L}_{reg},
$$

- Pseudo exemplar pairs:
  - VGG feature matching
- Real exemplar pairs:
  - Perceptual loss
  - Contextual loss
- Domain alignment loss
  - Domain l1
- Correspondence regularization
  - Cyclic warping loss
- Adversarial loss:
  - hinge loss
  - Discriminator feature matching

Section

Experiments

Quantitative Comparison

Table 1: **Image quality comparison.** Lower FID or SWD score indicates better image quality. The best scores are highlighted.

	ADE20k		ADE20k-*outdoor*		CelebA-HQ		DeepFashion	
	FID	SWD	FID	SWD	FID	SWD	FID	SWD
Pix2pixHD	81.8	35.7	97.8	34.5	62.7	43.3	25.2	**16.4**
SPADE	33.9	19.7	63.3	21.9	31.5	26.9	36.2	27.8
MUNIT	129.3	97.8	168.2	126.3	56.8	40.8	74.0	46.2
SIMS	N/A	N/A	67.7	27.2	N/A	N/A	N/A	N/A
EGSC-IT	168.3	94.4	210.0	104.9	29.5	23.8	29.0	39.1
Ours	**26.4**	**10.5**	**42.4**	**11.5**	**14.3**	**15.2**	**14.4**	17.2

Quantitative Comparison

Table 2: **Comparison of semantic consistency.** The best scores are highlighted.

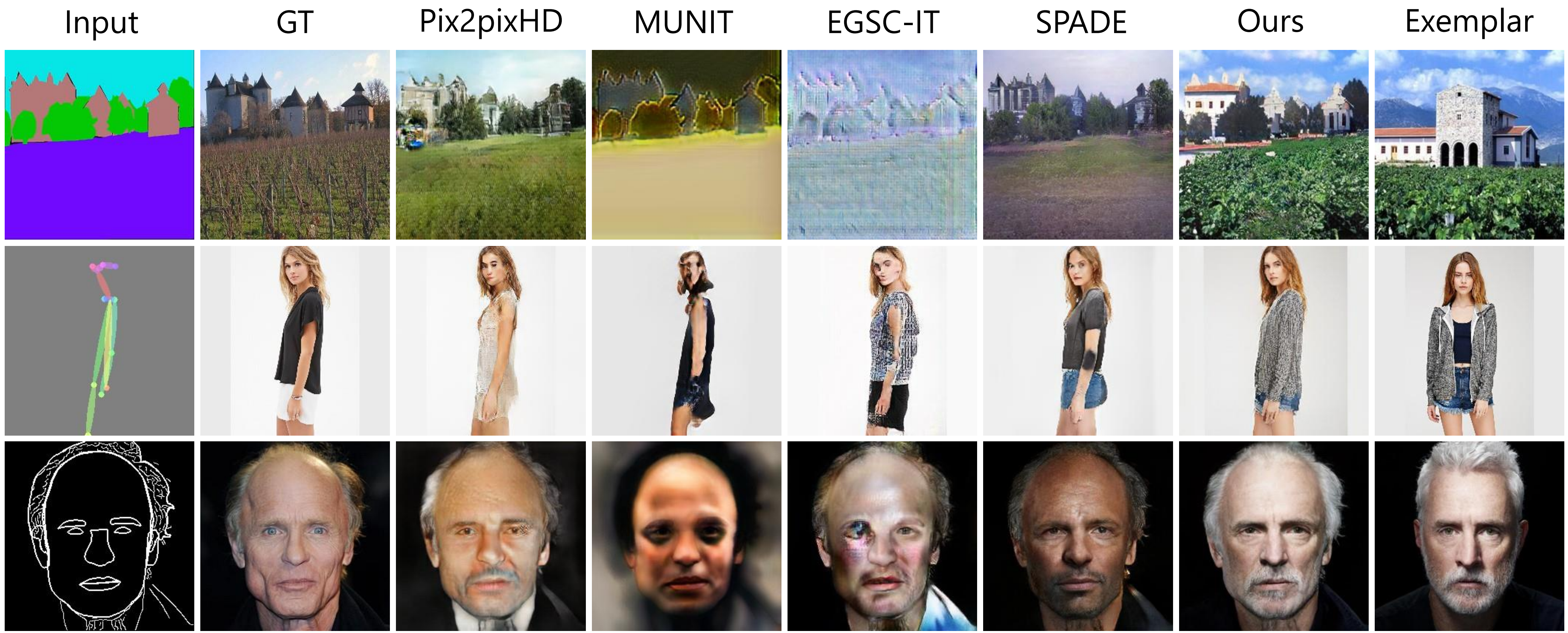
	ADE20k	ADE20k-*outdoor*	CelebA-HQ	DeepFashion
Pix2pixHD	0.833	0.848	0.914	0.943
SPADE	0.856	0.867	0.922	0.936
MUNIT	0.723	0.704	0.848	0.910
SIMS	N/A	0.822	N/A	N/A
EGSC-IT	0.734	0.723	0.915	0.942
Ours	**0.862**	**0.873**	**0.949**	**0.968**

Quantitative Comparison

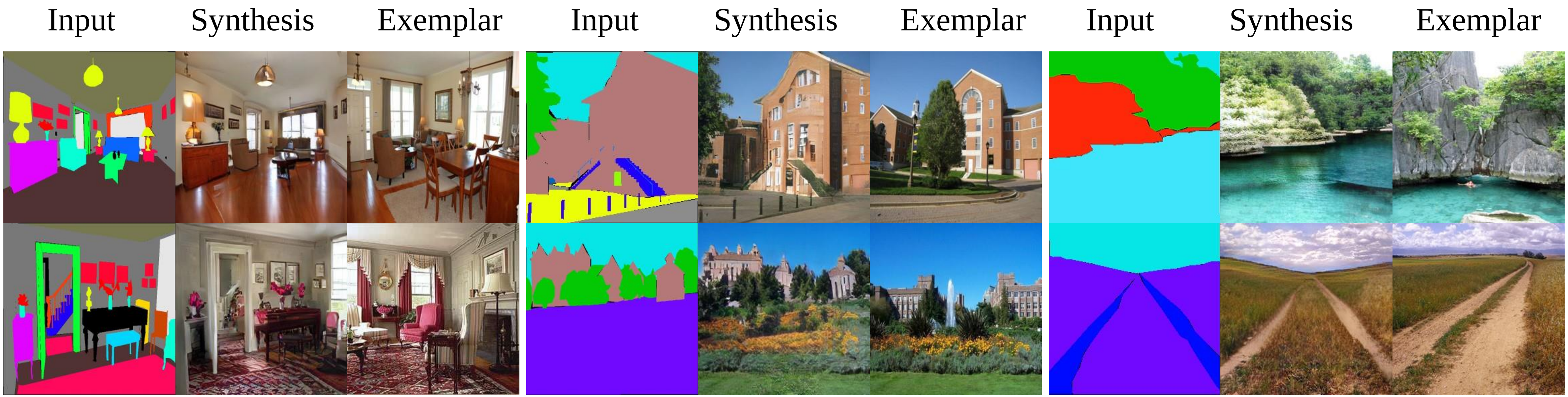
Table 3: **Comparison of style relevance.** A higher score indicates a higher appearance similarity relative to the exemplar. The best scores are highlighted.

	ADE20k		CelebA-HQ		DeepFashion	
	Color	Texture	Color	Texture	Color	Texture
SPADE	0.874	0.892	0.955	0.927	0.943	0.904
MUNIT	0.745	0.782	0.939	0.884	0.893	0.861
EGSC-IT	0.781	0.839	0.965	0.942	0.945	0.916
Ours	**0.962**	**0.941**	**0.977**	**0.958**	**0.982**	**0.958**

Comparison

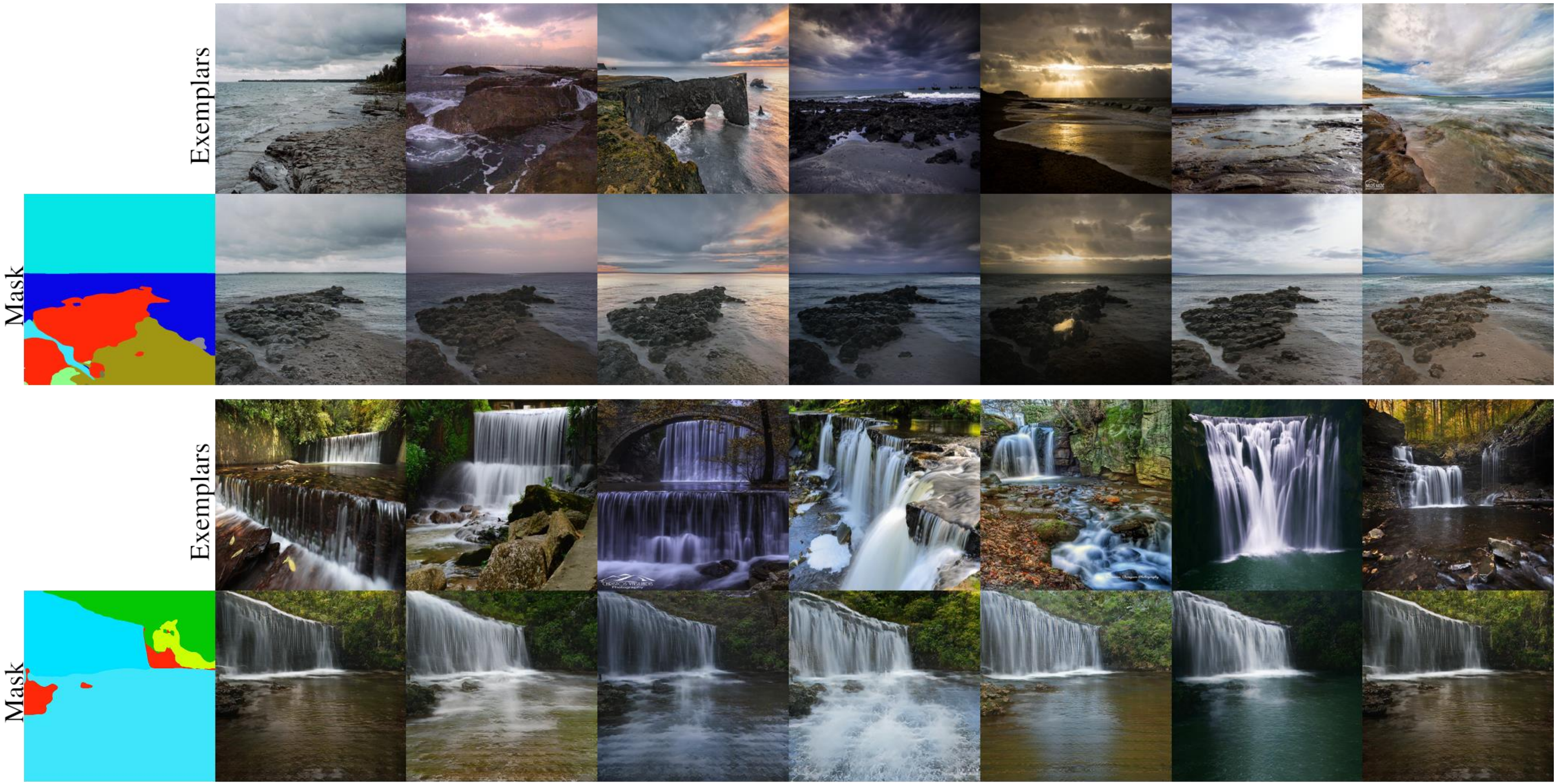


Mask-to-image synthesis: ADE20k



Mask-to-image synthesis: ADE20k

Mask-to-face: CelebA-HQ


Exemplars
Mask
Exemplars
Mask

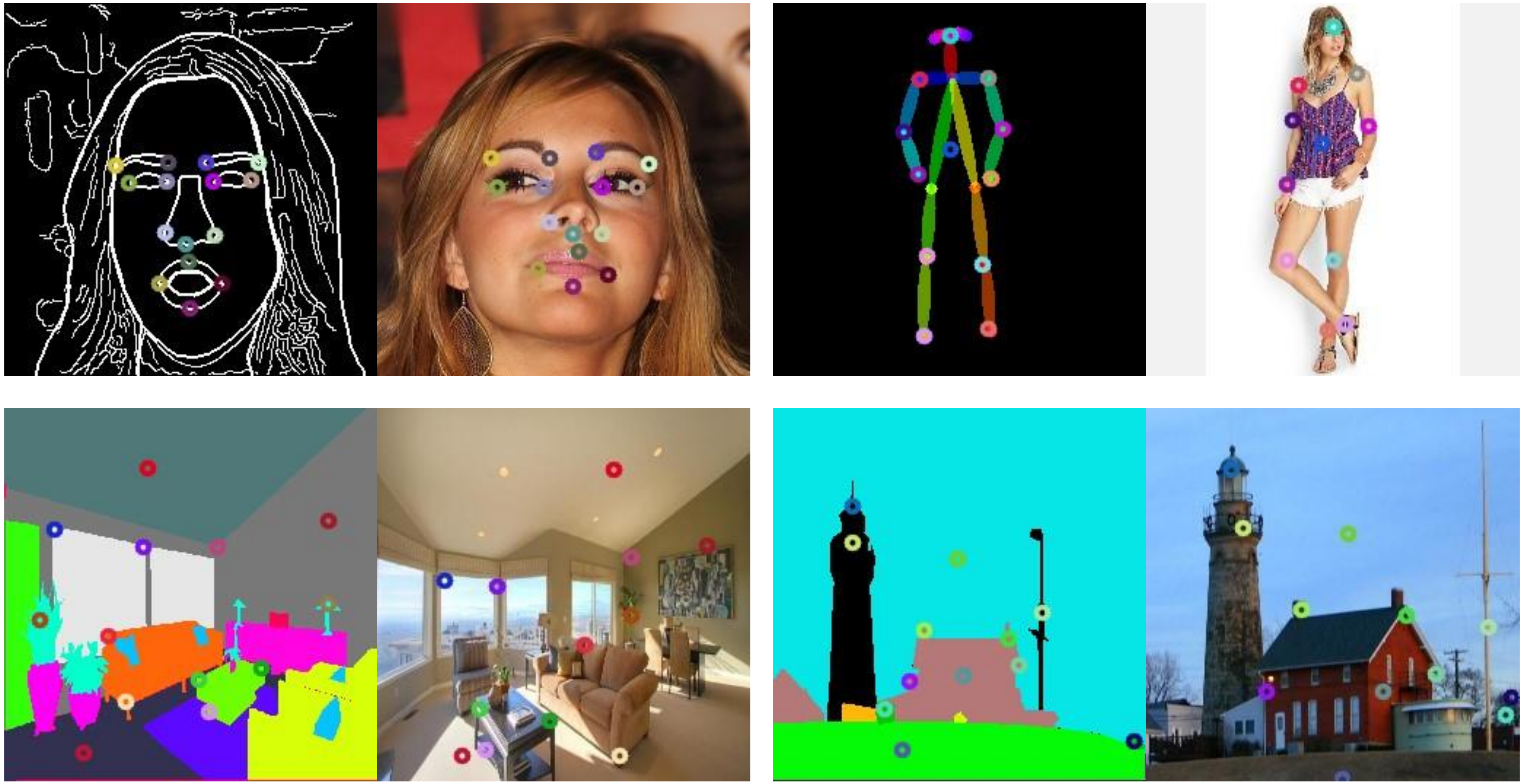
Edge-to-image synthesis: CelebA-HQ

Pose-to-image synthesis: DeepFashion

Section

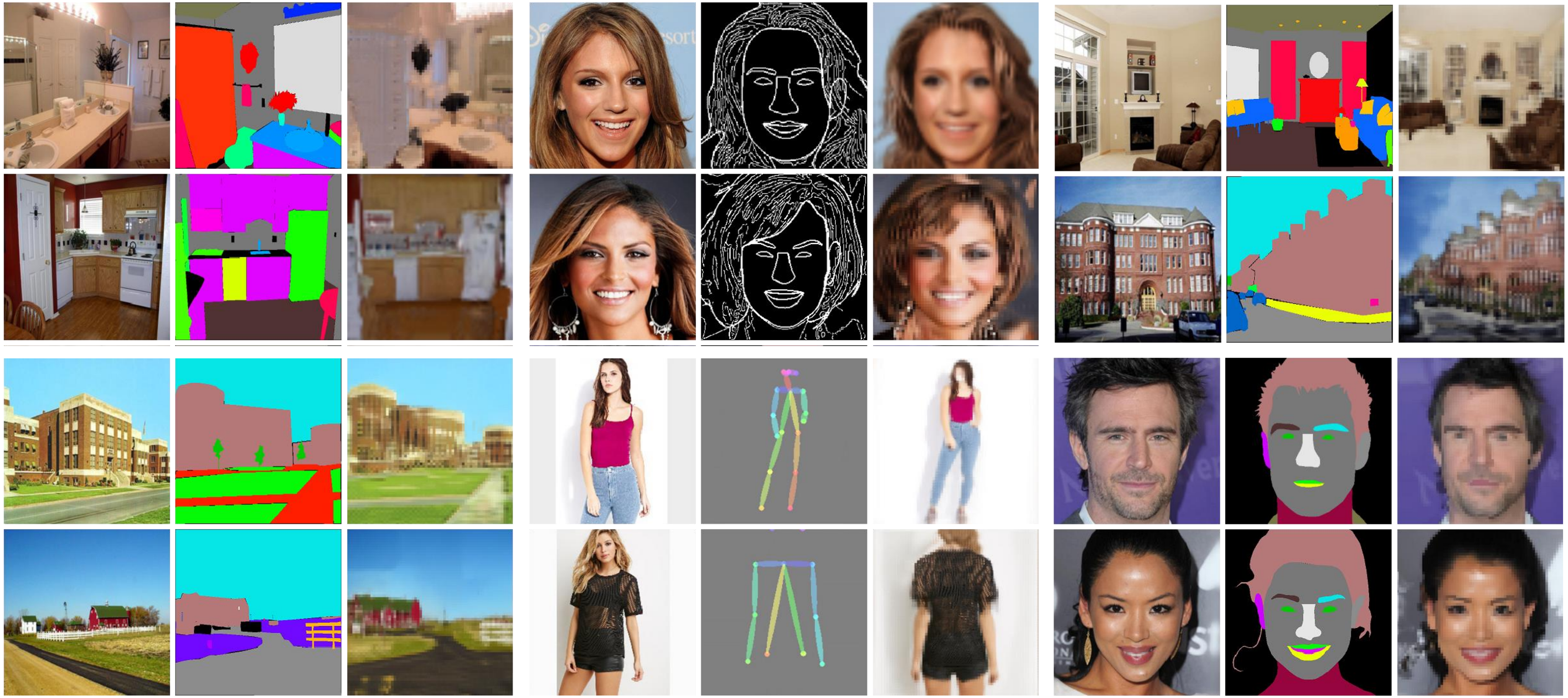
Application

Sparse Correspondence

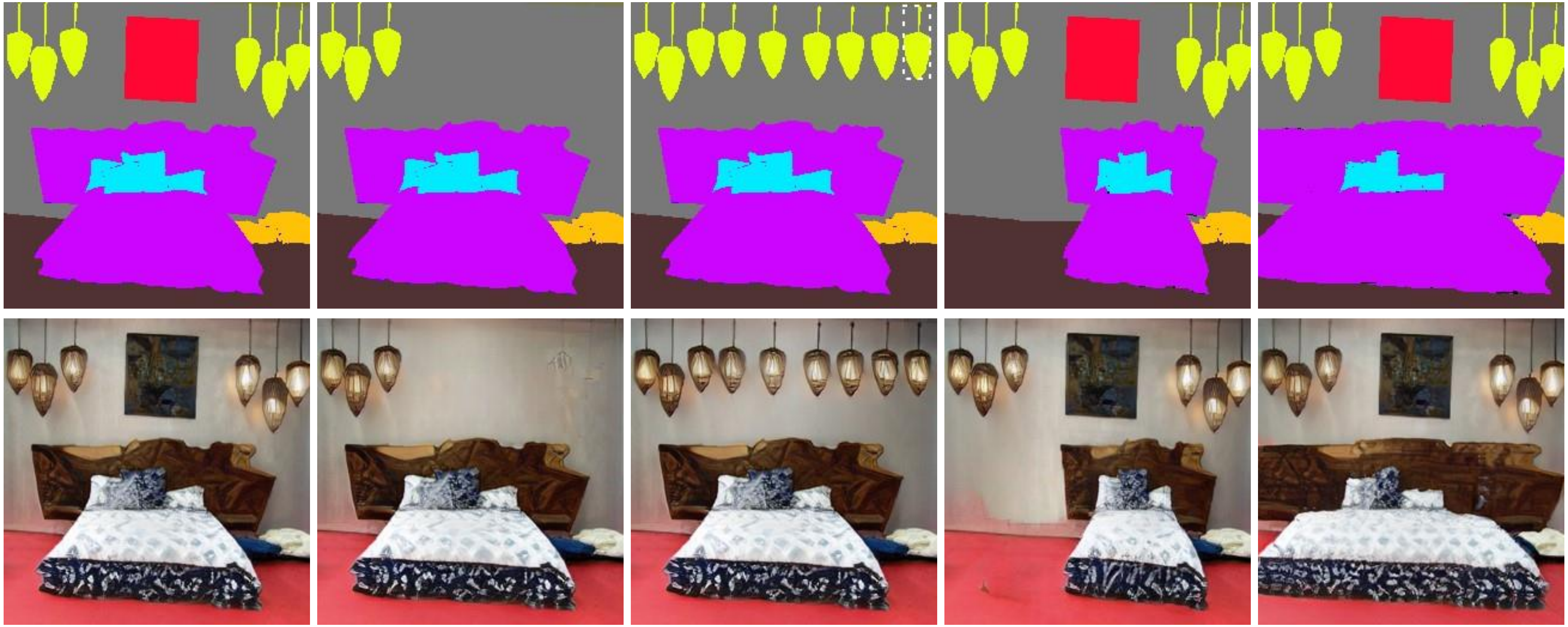


Sparse correspondence of different domains.

Dense Correspondence



Application: Image editing (Demo)



Application: Makeup transfer

Given a portrait along with makeup edits (1st column), we can transfer the makeup to other portraits by matching the semantic correspondence

Application: Makeup transfer

Figure 10: **Makeup transfer.** Given a portrait along with makeup edits (1st column), we can transfer the makeup to other portraits by matching the semantic correspondence.

Limitation

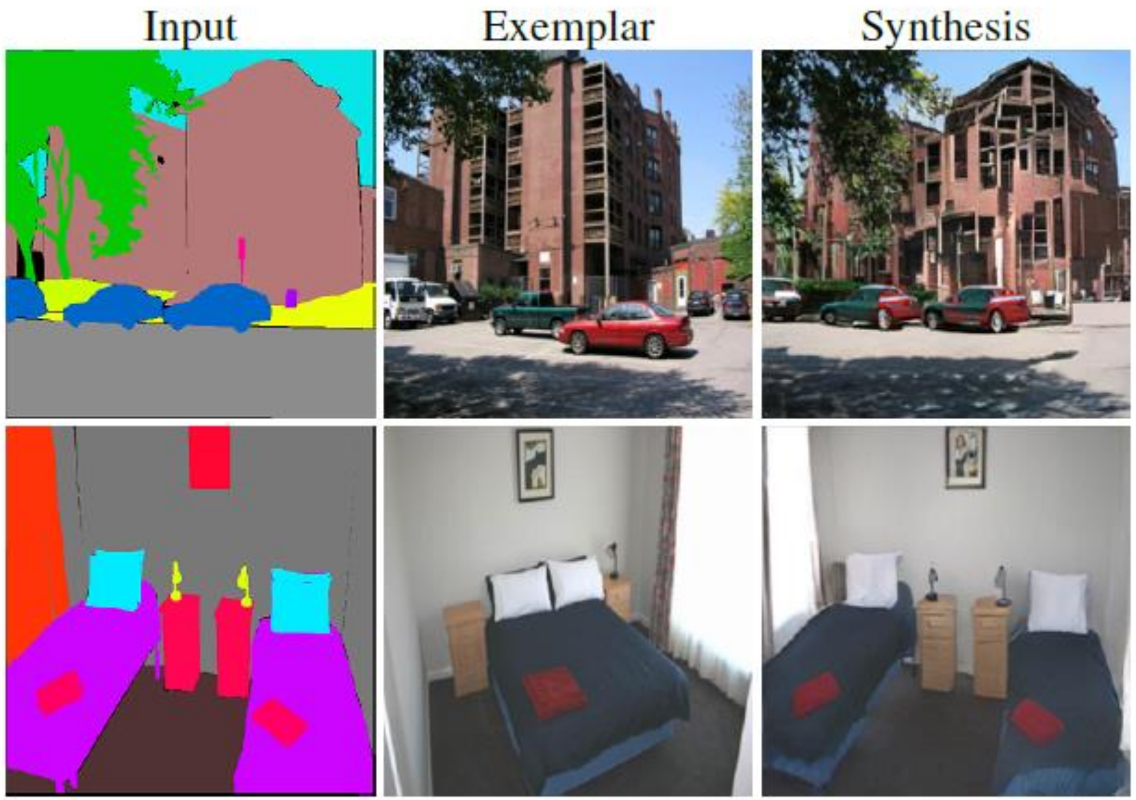


Figure 12: **Limitation.** Our method may produce mixed color artifact due the one-to-many mapping (1st row). Besides, the multiple instances (pillows in the figure) may use the same style in the cases of many-to-one mapping (2nd row).

Thank you!