Supplementary Material for
Cross-domain Correspondence Learning for Exemplar-based Image Translation

1. Additional Qualitative Results

Mask-to-image We perform mask-to-image synthesis on three datasets — ADE20k, CelebA-HQ and Flickr dataset, and we show their results in Figure 1-3 respectively.

Figure 1: **Our results of mask-to-image synthesis (ADE20k dataset).** In each group, the first row shows exemplars, and the second row shows the segmentation masks along with our results.

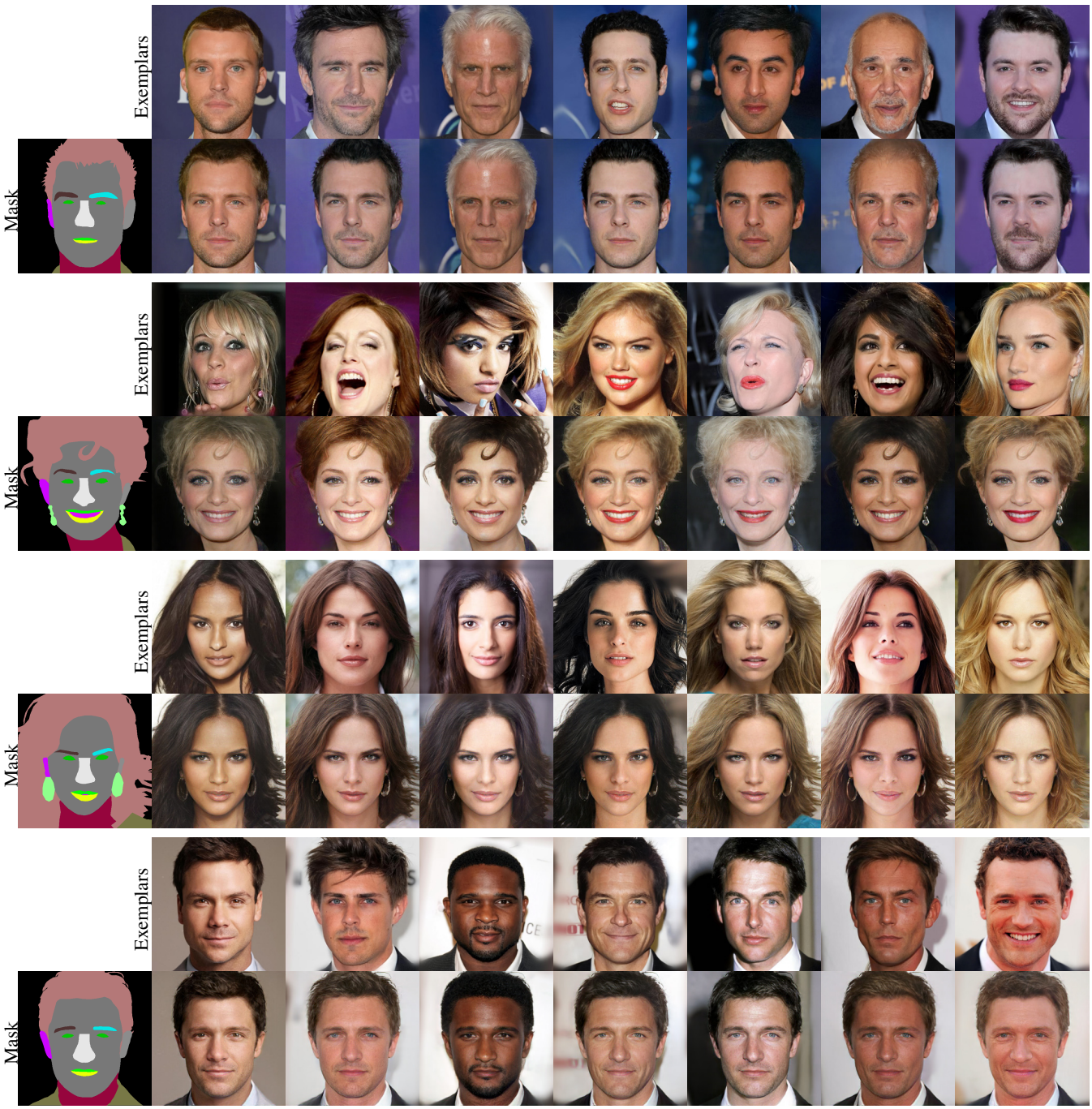

Figure 2: **Our results of mask-to-image synthesis (CelebAHQ dataset).** In each group, the first row shows exemplars, and the second row shows the segmentation masks along with our results.

Figure 3: **Our results of mask-to-image synthesis (Flickr dataset).** In each group, the first row shows exemplars, and the second row shows the segmentation masks along with our results.

Edge-to-face Figure 4 shows additional results of edge-to-face synthesis on CelebA-HQ dataset.

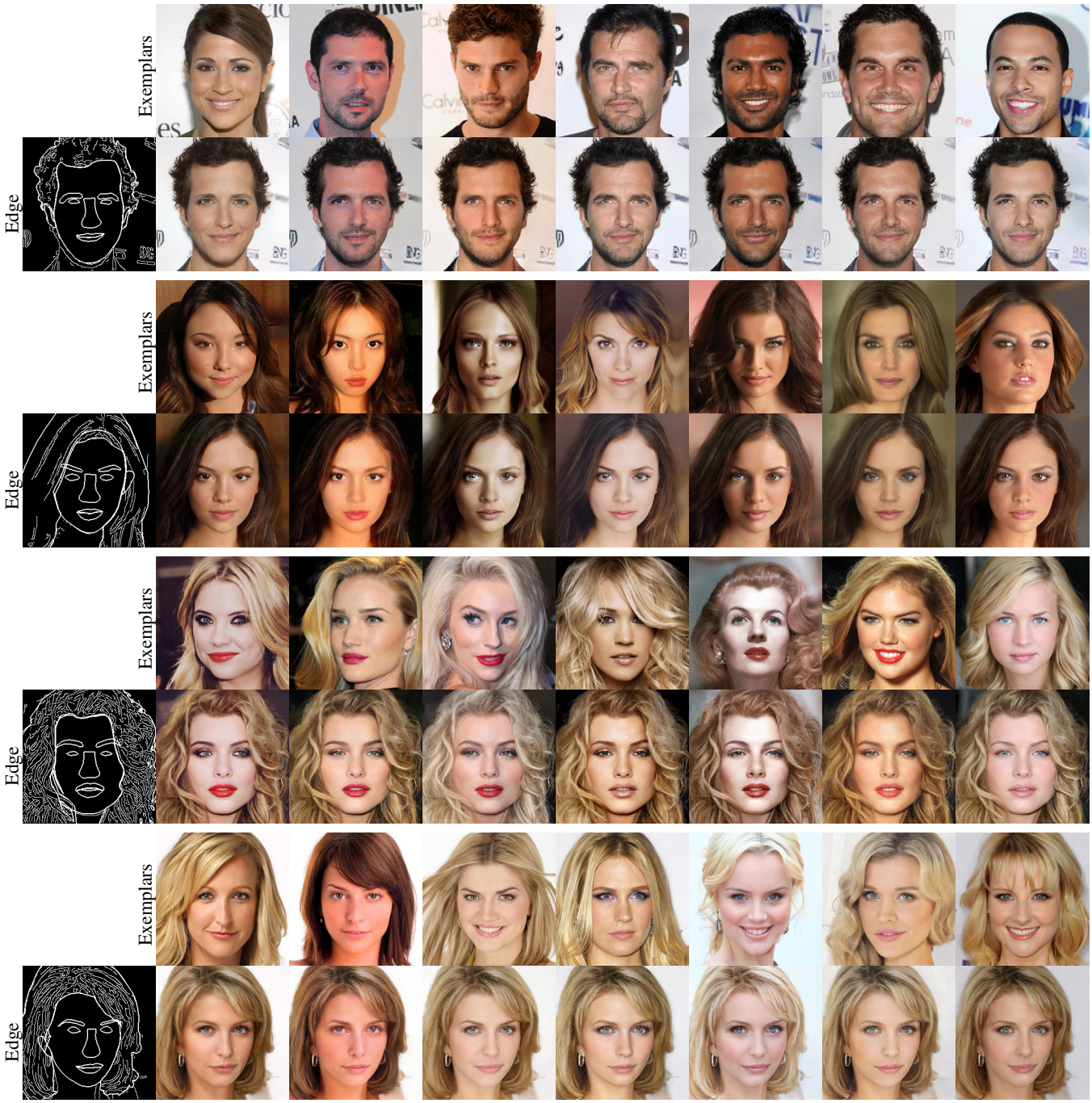

Figure 4: **Our results of edge-to-face synthesis (CelebA-HQ dataset).** In each group, the first row shows exemplars, and the second row shows the edge maps along with our results.

Pose-to-body Figure 5 shows more pose synthesis results on DeepFashion dataset.

Figure 5: **Our results of pose to image synthesis (DeepFashion dataset).** In each group, the first row shows exemplars, and the second row shows the pose images along with our results.

2. Additional Results of Dense Correspondence

The proposed *CoCosNet* is able to establish the dense correspondence between different domains. Figure 6 shows the dense warping results from domain B to domain A according to the correspondence ($r_{y\to x}$ in Equation 4).

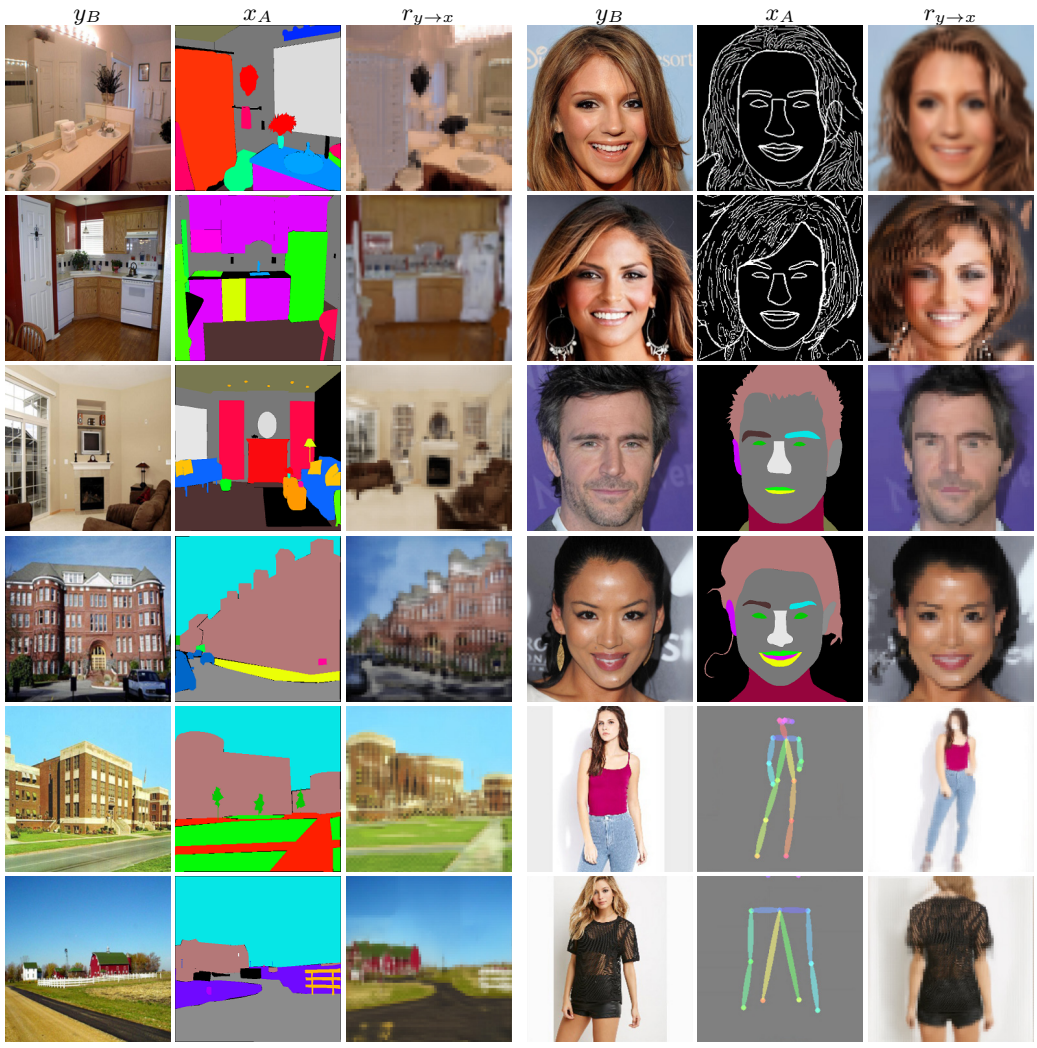

Figure 6: **Warping according to the dense correspondence.** The warped image $r_{y\to x}$ is semantically aligned as the image in domain A.

3. Additional Ablation Studies

Positional Normalization In the translation sub-network, we empirically find the normalization that computes the statistics at each spatial position better preserves the structure information synthesized in prior layers. Such positional normalization significantly improves the lower bound of our approach. We show the *worst case* result of ADE20k dataset in Figure 7, where the normalization helps produce vibrant image even when the correspondence is hard to be established in the complex scene.

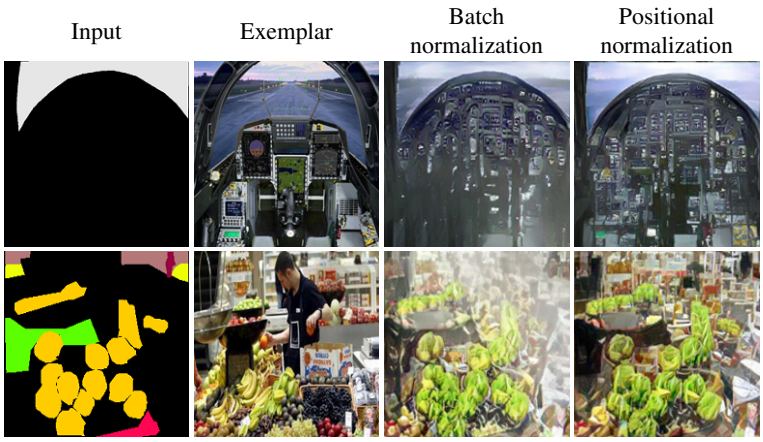

Figure 7: **Positional Normalization vs. Batch Normalization.** The positional normalization significantly improves the *lower bound* of the translation image quality.

Feature normalization for correspondence Note that we normalize the features before computing the correlation matrix. Likewise, we propose to calculate the statistics along the channel dimension while keeping the spatial size as the feature maps. This helps transfer the fine structures in the exemplar. As shown in Figure 8, the channel-wise normalization betters maintains the window structures in the ultimate output.

Figure 8: **Channel-wise normalization during correspondence.** The channel-wise normalization helps transfer the window structures from the exemplar image.

4. Additional Application Results

Image editing We show another example of the semantic image editing in Figure 9, where we manipulate the instances in the image by modifying their segmentation masks.

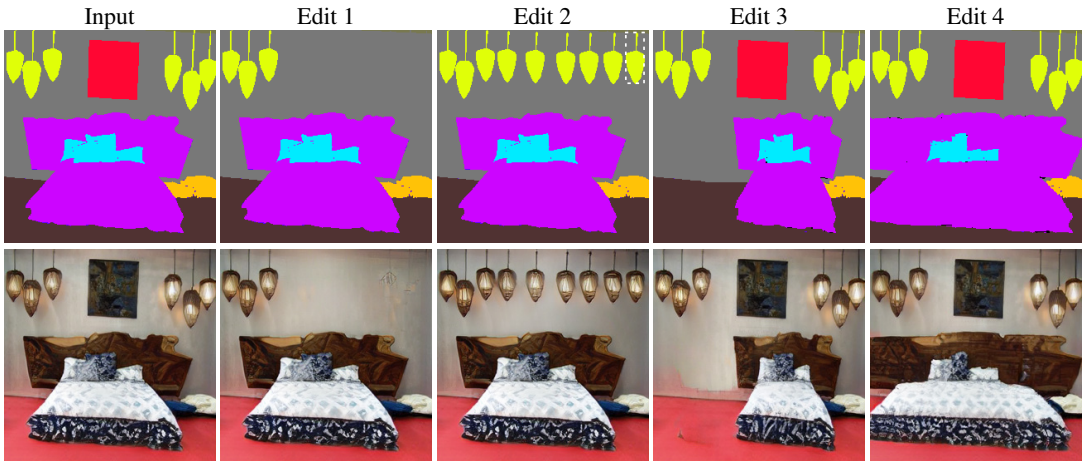

Figure 9: **Image editing.** Giving the original input image along with segmentation mask (1st column), we manipulate the image by changing its semantic layout (2nd-5th columns).

Makeup transfer Thanks to the dense semantic correspondence, we can transfer the makeup brushes to a batch of portraits. Figure 10 gives more supplementary results.

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.

5. Implementation Details

The detailed architecture of *CoCosNet* is shown in Table 1, with the naming convention as the CycleGAN.

Cross-domain correspondence network Two domain adaptors without weight sharing are used to adapt the input image and the exemplar to a shared domain S. The domain adaptors comprise several Conv-InstanceNorm-LeakReLU blocks and the spatial size of features in S is 64×64. Once the intermediate domain S is found, a shared adaptive feature block further transforms the features from two branches to the representation suitable for correspondence. The correlation layer computes pairwise affinity values between 4096×1 normalized features vectors. We downscale the exemplar image to 64×64 to fit the size of correlation matrix, and thus obtain the warped image on this scale. We use synchronous batch normalization within this sub-network.

Translation network The translation network generates the final output based on the style of the warped exemplar. We encode the exemplar style through two convolutional layers, which outputs α_i and β_i to modulate the normalization layer in the generator network. We have seven such style encoder, each responsible for modulating an individual normalization layer. The generator consists of seven normalization layer, which progressively utilizes the style code to synthesize the final output. The generator also employs a nonlocal block so that a larger receptive field can be utilized to enhance the global structural consistency. We use positional normalization within this sub-network.

Warm-up strategy For the most challenging ADE20k dataset, a mask warm strategy is used. At the beginning of the training, we explicitly provide the segmentation mask for the domain adaptors, and employ cross-entropy loss to encourage that the masks are correctly aligned after dense warping. Such warm-up helps speed up the convergence of the correspondence network and improve the correspondence accuracy. After training 80 epochs, we replace the segmentation masks with Gaussian noise. We just use the segmentation mask for warm up and there is no need to provide the masks during inference.

Table 1: **The architecture of *CoCosNet*.** k3s1 indicates the convolutional layer with kernel size 3 and stride 1. The ith style encoder outputs features with dimensions matching the ith Resblock in the generator.

Sub-network	Module	Layers in the module	Output shape (H×W×C)
Correspondence Network	Domain adaptor×2	Conv2d / k3s1	256×256×64
		Conv2d / k4s2	128×128×128
		Conv2d / k3s1	128×128×256
		Conv2d / k4s2	64×64×512
		Conv2d / k3s1	64×64×512
		Resblock×3 / k3s1	64×64×256
	Adaptive feature block	Resblock×4	64×64×256
		Conv2d / k1s1	64×64×256
	Correspondence	Correlation&warping	64×64×3
Translation Network	Style encoder×7	Bilinear interpolation	$h^i \times w^i \times 3$
		Conv2d / k3s1	$h^i \times w^i \times 128$
		Conv2d / k3s1	$h^i \times w^i \times c^i$
	Generator	Conv2d / k3s1	8×8×1024
		Resblock×5	128×128×256
		Nonlocal	128×128×256
		Resblock×2	256×256×64
		Conv2d / k3s1	256×256×3

6. Detailed User Study Results

Figure 11 shows the detailed results of user study. In ADE20k, there are 67.3% and 91.9% users respectively that prefer the image quality and style relevance for our method. Regarding edge-to-face translation on CelebA-HQ, 91.3% users prefer our image quality while 90.6% users believes our method most resembles the exemplar. For pose synthesis on DeepFashion dataset, 90.6% and 98.8% users prefer our results according to the image quality and the style resemblance respectively.

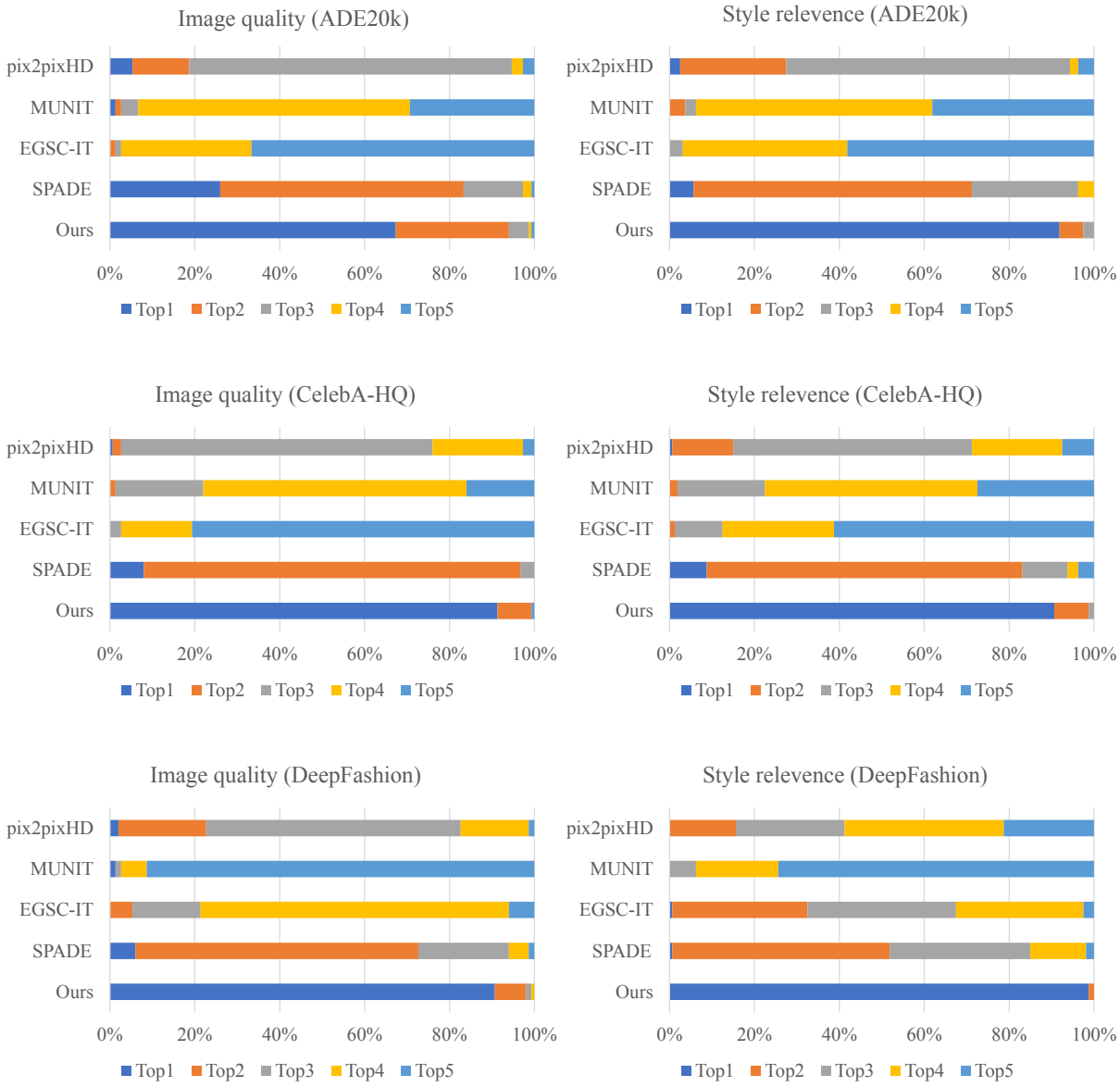

Figure 11: **Detailed user study results for ADE20k, CelebA-HQ and DeepFashion dataset.**

7. Multimodal results for Flickr dataset

Similar to the practice in [1], we collect 56,568 landscape images from Flickr. The semantic segmentation masks are computed using a pre-trained UPerNet101 [2] network. By feeding different exemplar, our method supports multimodal landscape synthesis. Figure 12 shows highly realistic landscape results using the images in Flicker dataset.

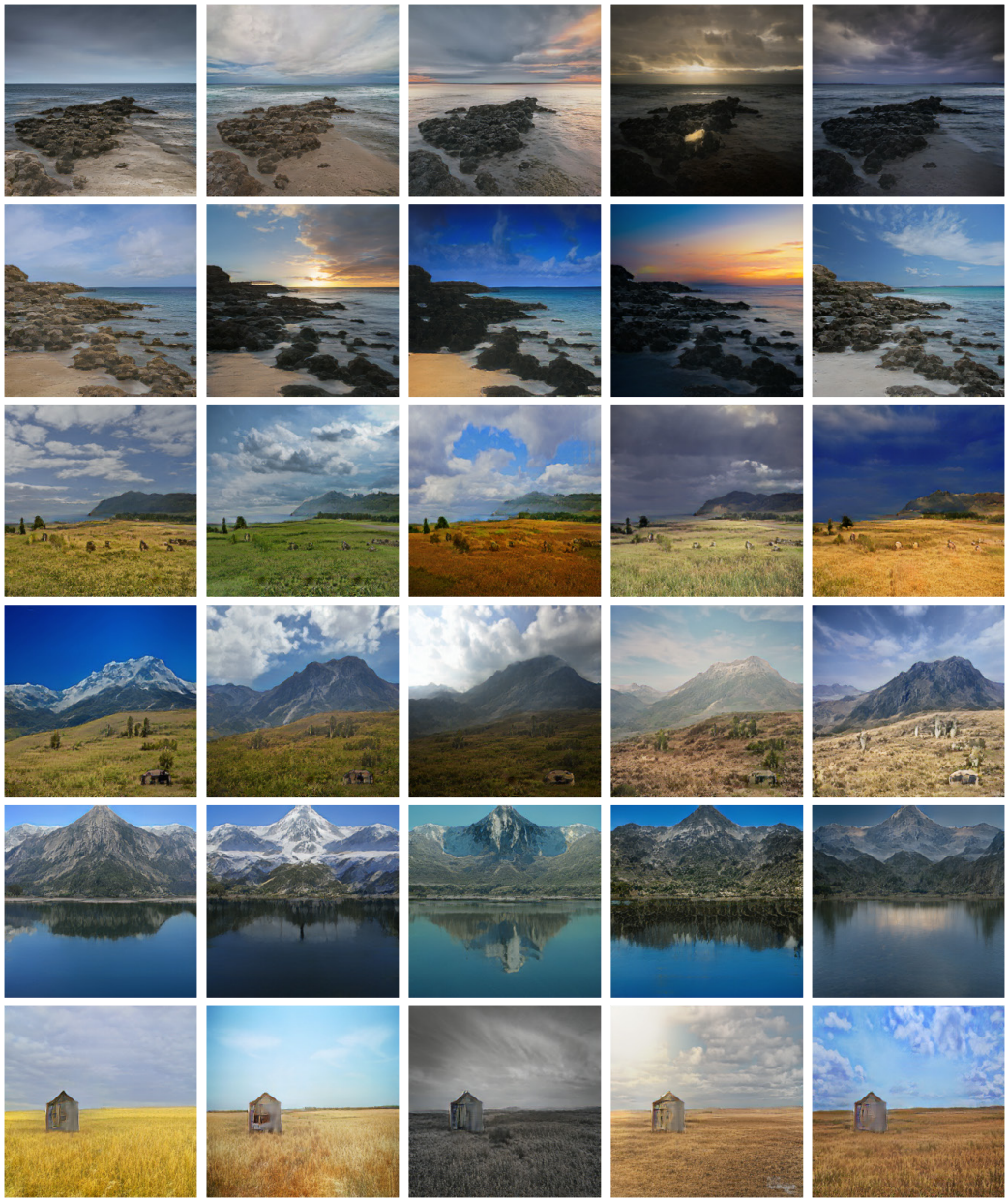

Figure 12: **Multimodal results of Flickr dataset.** We only present the final synthesis results here.

8. Limitation

As an exemplar-based approach, our method may not produce satisfactory results due to one-to-many and many-to-one mappings as shown in Figure 13. We leave further research tackling these issues as future work.

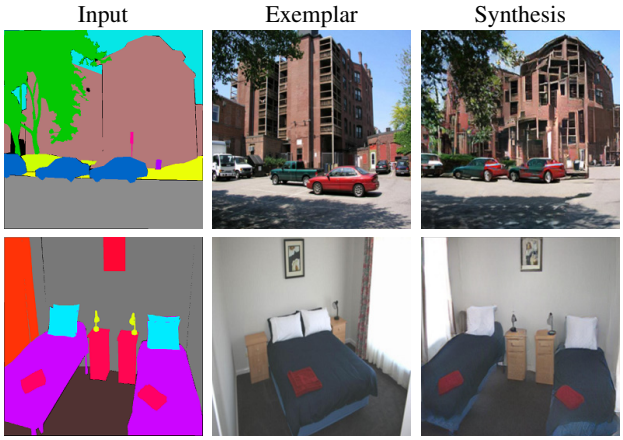

Figure 13: **Limitation.** Our method may produce mixed color artifact due the one-to-many mapping (1st row). Besides, the multiple instances (pillows in the figure) may use the same style in the cases of many-to-one mapping (2nd row).

Another limitation is that the computation of the correlation matrix takes tremendous GPU memory, which makes our method hardly scale for high resolution images. We leave the solve of this issue in future work.

References

[1] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346. 11

[2] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434. 11