DynaST: Dynamic Sparse Transformer for Exemplar-Guided Image Generation – *Supplementary Materials* –

Songhua Liu, Jingwen Ye, Sucheng Ren, and Xinchao Wang*

National University of Singapore
{songhua.liu,suchengren}@u.nus.edu, {jingweny,xinchao}@nus.edu.sg

In this document, we provide the additional details of the position embedding part and more results of the proposed DynaST model.

1 Position Embedding

We concatenate positional embedding to F_{tgt} and F_{ref} before computing their attention scores:

$$F_{tgt}^{i,pos} = [F_{tgt}^{i}, \mathrm{pos_i}], F_{ref}^{i,pos} = [F_{ref}^{i}, \mathrm{pos_i}], \tag{1}$$

where $0 \leq i < M$. The position embedding for the coarsest scale pos_{M-1} is a learnable tensor with the same spatial dimension as F^{M-1}. For upper levels ($0 \leq i < M-1$), position encoding $\mathrm{pos_i}$ is therefore generated with learnable upsample-convolution-nonlinear blocks based on $\mathrm{pos_{i+1}}$:

$$\mathrm{pos_i} = \mathrm{LReLU}(\mathrm{Conv}(\mathrm{Up}(\mathrm{pos_{i+1}}))), \tag{2}$$

where LReLU denotes leaky ReLU activation function.

2 More Results

Supervised Tasks. We provide more examples to better demonstrate advantages of our DynaST over state-of-the-art exemplar-guided image generation methods, including UNITE [6], CoCosNet [7], and CoCosNet-v2 [9]. On one hand, as shown in Fig. 1 Left, DynaST generates better local details compared with other methods in pose-guided person image generation task on the *DeepFashion* dataset, *e.g.*, cloth appearances in the 1st, 5th, 6th, and 7th rows, hats in the 2nd, 8th, and 9th rows, and cloth textures in the 4th and 10th rows. In particular, when there is a scale variance between exemplar and target images like the 3rd row, DynaST yields the best cloth-appearances and -textures restoration results, due to its dynamic attention mechanism. On the other hand, in edge-based face synthesis on the *CelebA-HQ* dataset shown in Fig. 1 Right, thanks to the construction of full-resolution matching, our results best capture

* Corresponding author.

local face details and global styles, *e.g.*, face details in the 1st, 3rd, and 5th rows, hand details in the 2nd row, beards in the 4th, 7th, and 9th rows, hair in the 8th row, mouth color in the 10th row, and global color patterns in the 6th row.
Undistorted Image Style Transfer. In Fig. 2, we show more comparisons with other state-of-the-art undistorted style transfer techniques, including LST [3], MST [8], WCT2 [5], MCCNet [1], AdaAttN [4], and MAST [2]. It turns out that the full-resolution matching mechanism in DynaST significantly improves the preservation of local details, such as the 1st, 3rd, 5th, and 6th rows, which in turn enables DynaST to better handle more complicated scenes like the 7th, 8th, and 9th rows. It also performs well when there are extreme textures in the style images, as shown in the 4th row. Meanwhile, the migration of global styles also outperforms those from previous approaches, *e.g.*, the 2nd and 10th rows.
Target-Exemplar Pairs. To further illustrate the performance of our proposed DynaST, we show results under pairwise input semantics and exemplar images in Fig. 3, Fig. 4, and Fig. 5 for the three tasks respectively. The images are randomly selected, which demonstrates the robustness of DynaST to different types of input and different scale variances between input and exemplar images.

References

1. Deng, Y., Tang, F., Dong, W., Huang, H., Ma, C., Xu, C.: Arbitrary video style transfer via multi-channel correlation. arXiv preprint arXiv:2009.08003 (2020)
2. Huo, J., Jin, S., Li, W., Wu, J., Lai, Y.K., Shi, Y., Gao, Y.: Manifold alignment for semantically aligned style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14861–14869 (2021)
3. Li, X., Liu, S., Kautz, J., Yang, M.H.: Learning linear transformations for fast image and video style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3809–3817 (2019)
4. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6649–6658 (2021)
5. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9036–9045 (2019)
6. Zhan, F., Yu, Y., Cui, K., Zhang, G., Lu, S., Pan, J., Zhang, C., Ma, F., Xie, X., Miao, C.: Unbalanced feature transport for exemplar-based image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15028–15038 (2021)
7. Zhang, P., Zhang, B., Chen, D., Yuan, L., Wen, F.: Cross-domain correspondence learning for exemplar-based image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5143–5153 (2020)
8. Zhang, Y., Fang, C., Wang, Y., Wang, Z., Lin, Z., Fu, Y., Yang, J.: Multimodal style transfer via graph cuts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5943–5951 (2019)
9. Zhou, X., Zhang, B., Zhang, T., Zhang, P., Bao, J., Chen, D., Zhang, Z., Wen, F.: Cocosnet v2: Full-resolution correspondence learning for image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11465–11475 (2021)

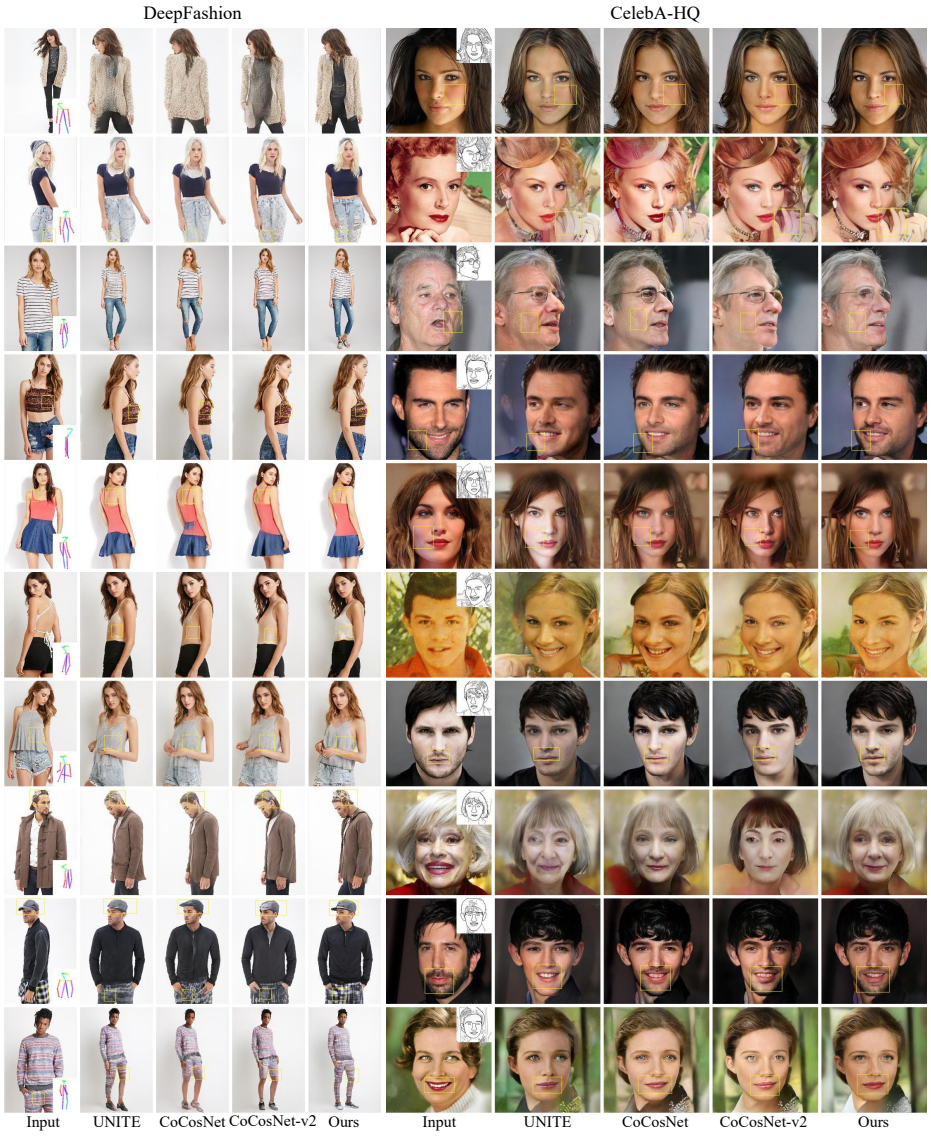

Fig. 1. More comparisons with state-of-the-art exemplar-guided image generation methods on two datasets.

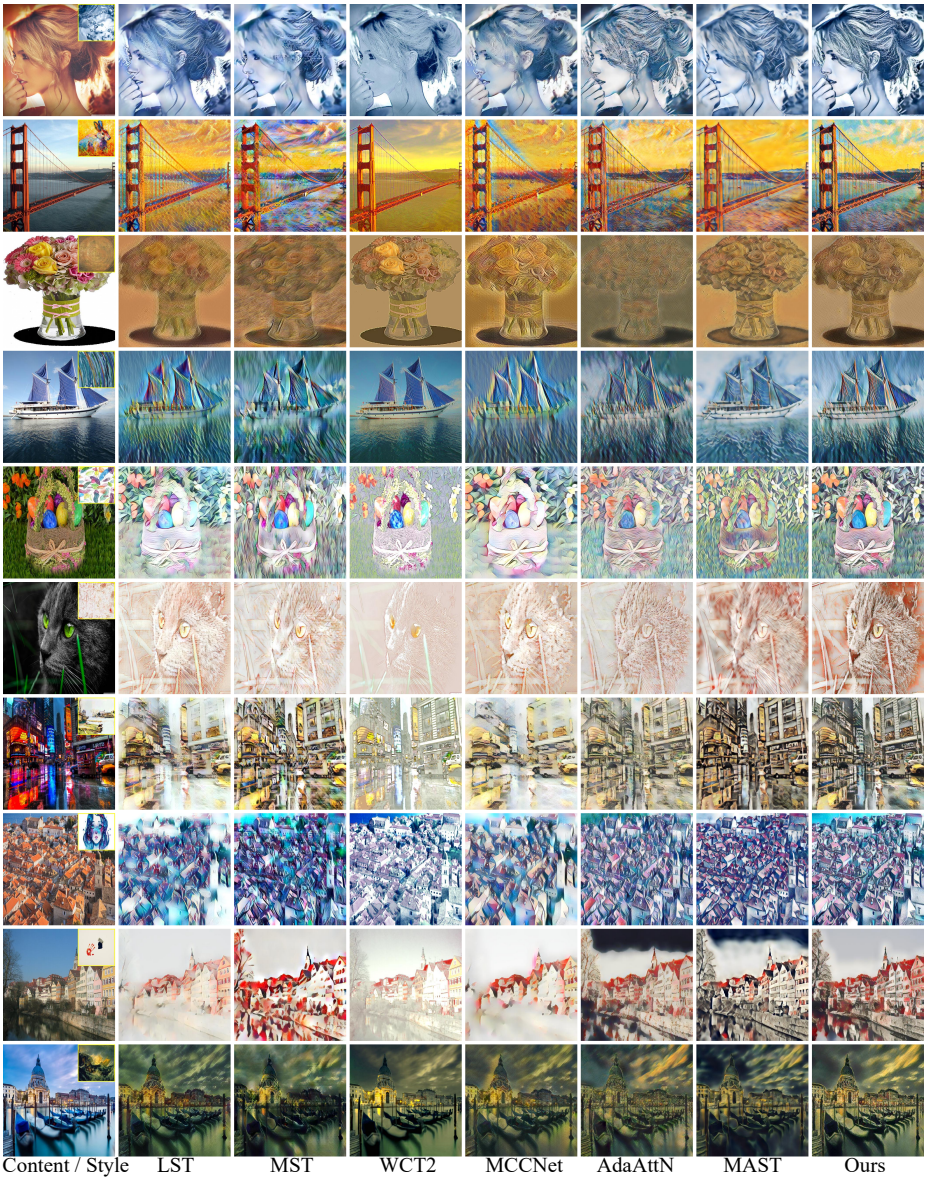

Fig. 2. More comparisons with state-of-the-art undistorted style transfer methods.

Fig. 3. More results by DynaST on pose-guided person image generation.

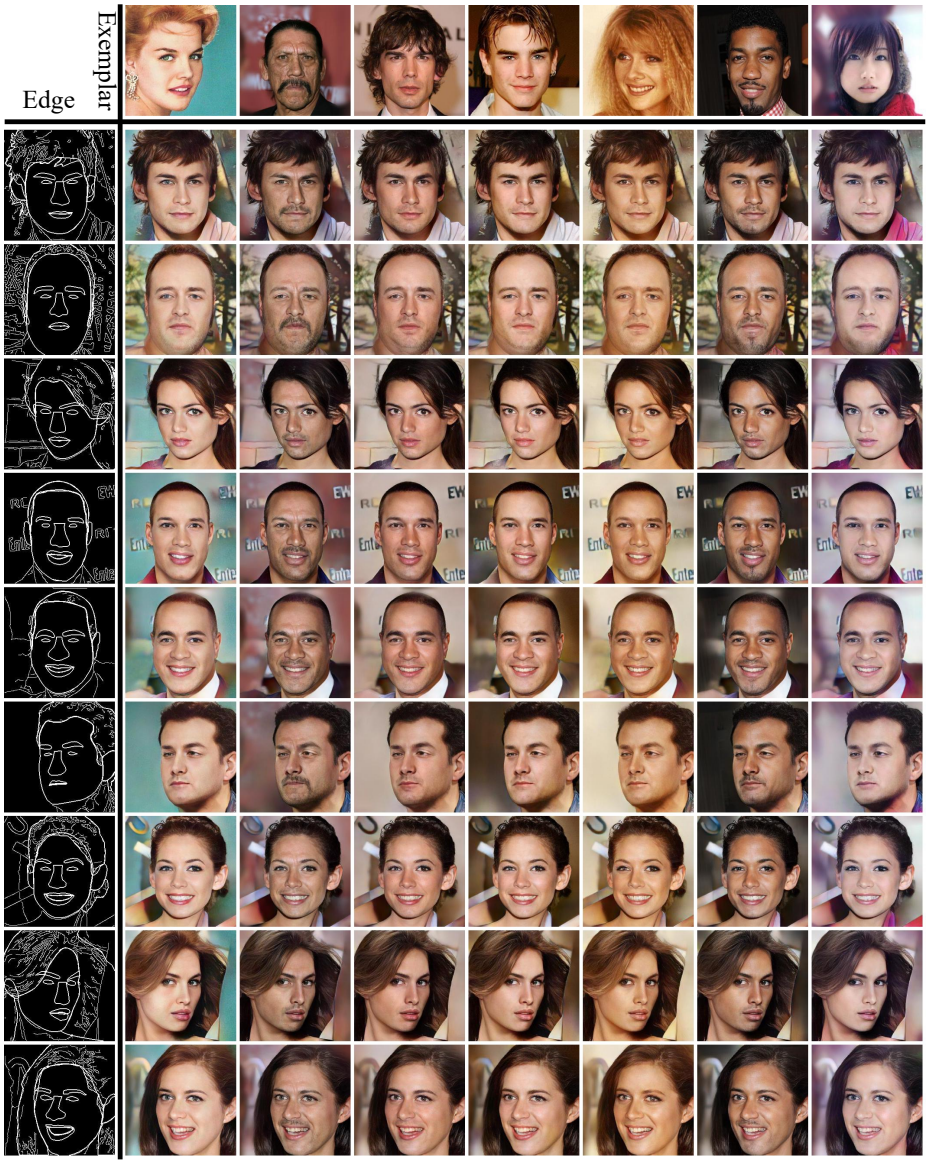

Fig. 4. More results by DynaST on edge-based face synthesis.

Fig. 5. More results by DynaST on undistorted image style transfer.