MixSyn: Compositional Image Synthesis with Fuzzy Masks and Style Fusion Supplementary Material

İlke Demir
Intel Labs
ilke.demir@intel.com

Umur Aybars Çiftçi
Binghamton University
uciftci@binghamton.edu

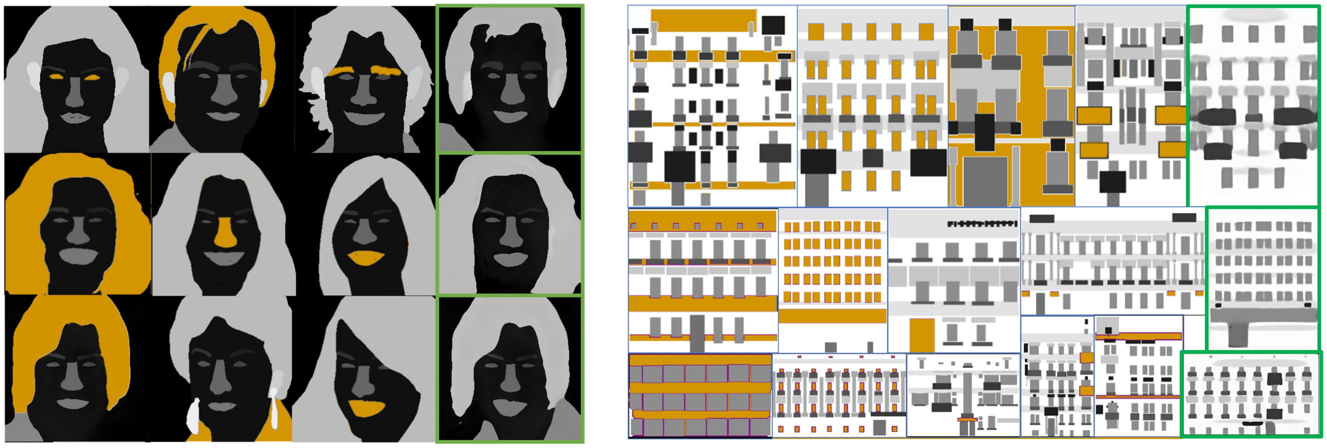

Figure 1. Compositions. Orange regions in each row (r_j^is) are given as input to MixSyn structure generator, to create coherent random compositions M'' (green) for three faces (left) and three facades (right).

A. Additional Compositions

We show more composition results parallel to Sec. 3.1 on faces and facades. Three different regions (colored in orange) are selected to learn face compositions (outlined in green), and four or five different regions are selected to learn facade compositions.

B. Network Architectures

In addition to Fig. 3 and Fig. 4 of the main paper, we document specific layers of our generative architectures, including output shapes and resampling/normalization layers per block. Tab. 1 (top) documents the structure encoders in Fig. 3a, Tab. 1 (mid) documents the structure decoder in Fig. 3b, Tab. 1 (bottom) documents the structure discriminator in Fig. 3c, Tab. 2 (top) documents the style encoder in Fig. 4a, Tab. 2 (middle) documents the style generator in Fig. 4c, Tab. 2 (bottom) documents the style discriminator in Fig. 4d, all using different configurations and combinations of MS block in Fig. 4e.

C. Region Reconstruction Scores

In addition to our full-image scores reported on the main paper, we would like to understand and evaluate the capabilities of MixSyn further, by analyzing its results per region. We compute *per region* FID, PSNR, RMSE, and SSIM scores on CelebAMask-HQ dataset (top) and on CMP Facade dataset (bottom) in Tab. 3, both for *known* composition and *approximated* image generation results. The results are normalized per occurrence, i.e., an image without a hat does not contribute to overall hat score. As we are comparing known/approximated instances which should replicate the originals, no alignment step is needed.

For face images, we observe that learning to generate hair styles consistently is still a bottleneck (FID=49.73), which we aim to get better at by training our network for more epochs when we have resources. For compositions, MixSyn does a good job for almost all regions, except hats (SSIM=0.78), meaning that it has an internal blurry understanding of where to put a hat, but the shape is not well-defined. We also observe that even though FID score of hat is low, other hat scores are relatively worse. We speculate

Layer	Resample	Norm	Output
Mask X	-	-	256x256x1
Conv1x1	-	-	256x256x16
MS	Avg Pool	IN	128x128x32
MS	Avg Pool	IN	64x64x64
MS	Avg Pool	IN	32x32x128
MS	Avg Pool	IN	16x16x128
MS	-	IN	16x16x128
MS	-	IN	16x16x128

Layer	Resample	Norm	Output
Concatenation	-	-	16x16x1920
MS	-	IN	16x16x1024
MS	-	IN	16x16x512
MS	-	IN	16x16x512
MS	-	IN	16x16x512
MS	Upsample	IN	32x32x256
MS	Upsample	IN	64x64x128
MS	Upsample	IN	128x128x64
MS	Upsample	IN	256x256x32
Conv1x1	-	-	256x256x1

Layer	Resample	Norm	Output
Mask X	-	-	256x256x1
Conv1x1	-	-	256x256x64
MS	Avg Pool	IN	128x128x128
MS	Avg Pool	IN	64x64x256
MS	Avg Pool	IN	32x32x512
MS	Avg Pool	IN	16x16x512
MS	Avg Pool	IN	8x8x512
MS	Avg Pool	IN	4x4x512
LReLU	-	-	4x4x512
Conv4x4	-	-	1x1x512
LReLU	-	-	1x1x512
Reshape	-	-	512
Linear	-	-	1

Table 1. Structure encoder architecture, per region type (top), decoder architecture (mid), and discriminator architecture (bottom).

that hats do not have features as specific as other regions, thus their inter-type similarity is naturally low.

For building images, we generally see smooth (low PSNR) and averaged (high RMSE) facades. This trend can be considered similar to what we observe in the skin region for faces. Regions with intricate details and small areas (e.g., pillars) and undersampled distributions (e.g., decorations) also tend to result in lower scores. For building compositions, the network learns rectangular regions as all the labels in the CMP facades dataset are axis-aligned boxes. In future, we hypothesize that a dataset with more exact masks (instead of approximate boxes) would significantly improve our image scores in the architecture domain.

Layer	Resample	Norm	Output
Image X	-	-	256x256x3
Conv1x1	-	-	256x256x16
MS	Avg Pool	IN	128x128x32
MS	Avg Pool	IN	64x64x64
MS	Avg Pool	IN	32x32x128
MS	Avg Pool	IN	16x16x128
MS	Avg Pool	IN	8x8x128
MS	Avg Pool	IN	4x4x128
LReLU	-	-	4x4x128
Conv4x4	-	-	1x1x128
LReLU	-	-	1x1x128
Linear	-	-	16

Layer	Resample	Norm	Output
Mask X	-	-	256x256x1
Conv1x1	-	-	256x256x32
MS	Avg Pool	IN	128x128x64
MS	Avg Pool	IN	64x64x128
MS	Avg Pool	IN	32x32x256
MS	Avg Pool	IN	16x16x256
MS	-	IN	16x16x256
MS	-	IN	16x16x256
MS	-	SEAN	16x16x256
MS	-	SEAN	16x16x256
MS	Upsample	SEAN	32x32x256
MS	Upsample	SEAN	64x64x128
MS	Upsample	SEAN	128x128x64
MS	Upsample	SEAN	256x256x32
Conv1x1	-	-	256x256x3

Layer	Resample	Norm	Output
Image X	-	-	256x256x3
Conv1x1	-	-	256x256x64
MS	Avg Pool	IN	128x128x128
MS	Avg Pool	IN	64x64x256
MS	Avg Pool	IN	32x32x512
MS	Avg Pool	IN	16x16x512
MS	Avg Pool	IN	8x8x512
MS	Avg Pool	IN	4x4x512
LReLU	-	-	4x4x512
Conv4x4	-	-	1x1x512
LReLU	-	-	1x1x512
Reshape	-	-	512
Linear	-	-	1

Table 2. Style encoder architecture, per region type (top), generator architecture (middle), and discriminator architecture (bottom).

	Image				Composition			
Region	FID	PSNR	RMSE	SSIM	FID	PSNR	RMSE	SSIM
Skin	13.070	22.199	5.210	0.862	79.318	22.964	2.600	0.920
EyeBrow$_l$	7.149	39.737	0.621	0.998	10.122	52.180	0.505	0.999
EyeBrow$_r$	10.783	39.288	0.623	0.998	22.697	45.704	0.597	0.997
Eye$_l$	7.388	38.496	0.488	0.998	21.498	47.149	0.363	0.998
Eye$_r$	10.470	39.661	0.481	0.998	35.418	43.671	0.332	0.997
Nose	18.162	35.046	1.358	0.992	8.636	40.329	0.501	0.996
Mouth	16.343	32.528	1.191	0.990	9.169	41.221	0.429	0.994
Cloth	20.401	27.151	1.861	0.959	10.526	30.926	0.926	0.983
Glasses	3.309	23.119	2.446	0.942	3.455	18.806	1.913	0.932
Necklace	2.624	38.410	0.609	0.996	3.602	31.877	0.644	0.991
Hair	49.739	20.004	5.599	0.717	28.054	22.039	2.004	0.935
Ear$_l$	19.053	31.076	1.136	0.988	12.253	30.386	1.021	0.992
Ear$_r$	15.852	32.474	1.105	0.989	14.513	27.982	1.018	0.991
Earring	12.753	34.942	0.814	0.990	5.967	32.586	0.884	0.993
Hat	8.975	15.728	4.767	0.777	6.795	8.079	3.538	0.784
All	14.405	31.324	1.887	0.946	18.135	33.060	1.152	0.967

	Image				Composition			
Region	FID	PSNR	RMSE	SSIM	FID	PSNR	RMSE	SSIM
Facade	12.537	21.560	5.389	0.840	9.375	11.763	6.385	0.467
Molding	5.663	48.267	2.447	0.962	6.105	18.375	3.290	0.799
Cornice	4.179	134.085	1.139	0.990	3.961	60.971	1.736	0.922
Pillar	1.883	203.441	0.712	0.995	2.143	112.121	1.203	0.942
Window	9.862	26.421	3.134	0.938	8.083	15.591	4.094	0.707
Door	3.757	164.023	0.672	0.993	3.668	92.503	1.046	0.977
Sill	4.636	117.907	1.055	0.993	3.930	46.061	1.606	0.927
Balcony	7.318	172.295	1.176	0.980	5.307	93.285	1.886	0.919
Decoration	3.622	168.235	0.780	0.993	3.214	88.142	1.217	0.954
All	15.750	17.816	7.545	0.591	20.145	21.928	3.917	0.894

Table 3. **Region Reconstruction Scores** on CelebAMask-HQ dataset (top) and on CMP Facades dataset (bottom), with known compositions and approximated images.

D. Region Similarity Scores

Following the reconstruction analysis on approximated images, we also evaluate MixSyn on *random* compositions and *random* images for region-based similarity (Tab. 4).

As regions transform during composition generation, final segments can have any shape, beclouding segment correspondences between the source and generated images. For comparison, we do not perform a full alignment, but we translate and scale the bounding boxes of each corresponding region and segment. This process introduces a slight downgrading effect on our random generation scores, however it still establishes a common ground to understand and compare the performance on different segments.

Comparing scores of random and known compositions, we observe that all scores are lower for random compositions, as expected. What is unexpected is that, comparing approximated and random images, MixSyn exploits this flexibility in compositions as a superpower to generate more realistic images, as all similarity metrics are higher for random images. Even the aforementioned hat region SSIM increases when the underlying mask is flexible. The only exception is the skin: When its composition is blurry, other regions mostly slide and occupy original skin pixels, thus its similarity scores get worse while RMSE gets better.

	Image			Composition		
Region	PSNR	RMSE	SSIM	PSNR	RMSE	SSIM
Skin	17.090	2.855	0.658	18.965	5.402	0.815
Brow(L)	42.332	0.477	0.993	38.139	0.653	0.997
Brow(R)	39.589	0.758	0.993	37.778	0.651	0.997
Eye(L)	43.639	0.251	0.997	37.918	0.491	0.997
Eye(R)	41.339	0.306	0.997	38.510	0.494	0.997
Nose	34.884	0.765	0.991	33.784	1.354	0.990
Mouth	37.252	0.483	0.991	32.245	1.180	0.990
Cloth	29.731	1.508	0.982	25.900	1.940	0.955
Glasses	17.258	2.615	0.921	22.460	2.528	0.938
Necklace	43.353	0.529	0.999	38.190	0.531	0.996
Hair	13.401	4.185	0.826	16.538	5.444	0.689
Ear(L)	29.309	1.114	0.990	29.039	1.133	0.986
Ear(R)	26.114	1.335	0.989	29.925	1.108	0.986
Earring	29.087	0.904	0.987	32.056	0.893	0.987
Hat	13.402	2.575	0.899	14.027	5.043	0.753
All Parts	30.519	1.377	0.947	29.698	1.923	0.938

Table 4. **Region Similarity Scores** on CelebAMask-HQ dataset, for random compositions and random images.

E. Cross-Dataset Reconstruction of Regions

Supporting our evaluation in Sec. 5.1 marked with (H), we list region reconstruction scores of the model trained on CelebAMask-HQ [3] and tested on Helen [4] in Tab. 5.

Originally, Helen dataset has fewer semantic classes than CelebAMask-HQ, so we create an internal mapping. We keep the mouth meta-class, as the combination of lips and mouth. For known compositions, overall scores increase with less region types. For approximated images, we observe similar trends for harder to generate inexact regions, i.e., hair and skin. However since there are no item types (necklace, hat, glasses, etc.), the network is more decisive about main types.

	Image			Composition		
Part	PSNR	RMS	SSIM	PSNR	RMS	SSIM
Skin	18.485	4.195	0.855	19.808	3.294	0.900
Br.$_l$	38.264	0.541	0.997	50.085	0.441	0.998
Br.$_r$	37.032	0.542	0.996	47.830	0.528	0.998
Eye$_l$	39.238	0.416	0.997	45.377	0.436	0.997
Eye$_r$	38.473	0.420	0.997	43.415	0.354	0.997
Nose	31.251	1.189	0.989	39.012	0.797	0.995
Mou.	32.322	0.941	0.991	43.310	0.440	0.996
Hair	22.009	3.476	0.866	36.006	1.098	0.989
All	32.134	1.465	0.961	40.605	0.923	0.984

Table 5. **Cross-Dataset Reconstruction Scores** per region, computed on Helen [4] and trained on CelebAMask-HQ [3], with known compositions and approximated images.

F. Quantitative Comparison

In Tab. 6, we report PSNR, RMSE, SSIM, and FID scores for 16 generated images in Fig. 7 of main text, with copy/paste (CP) and our (O) mask-image pairs. To clarify with an example, the cell at SPADE SSIM row and O/CP column reports the SSIM score of the image generated by SPADE using our mask and copy/paste image as the source. Each cell is spatially aligned with the image in Fig. 7 of the main text. For comparison, we concatenate results reported at the bottom of Fig. 8 of main text as the last column. For completeness, we report our generated image scores in the green cells as the first cell of the last column and the output of collage-based generation in the last cell of the last column, since SPADE and MaskGAN do not have component transfer applications or sequential generation.

In all four score types, our approach beats prior work in all combinations: given copy/paste mask and copy/paste image, or copy/paste mask and our image, or our mask and copy/paste image, as input. MixSyn also performs better than the sequential component transfer applications of SEAN and Mask Guided CGAN that utilize a base mask (rightmost cells).

As a validation test, we input our generated composition and our generated image to all approaches and compute their reconstruction score (column 4). Even with the expected output given as the input, their reconstruction scores are barely on par with our scores. Visually comparing such, they look like copied segments and do not blend as naturally as our images, because of the lack of interpretation of the underlying fuzzy mask.

Approach	Metric	CP/CP	CP/O	O/CP	O/O	Component transfer
SPADE [5]	PSNR	18.899	19.831	18.900	28.109	24.271 (ours)
	RMSE	7.232	7.160	7.229	2.587	2.782 (ours)
	SSIM	0.711	0.730	0.712	0.882	0.840 (ours)
	FID	15.772	22.118	15.650	14.442	13.125 (ours)
SEAN [6]	PSNR	19.741	22.029	20.471	30.112	21.995
	RMSE	7.168	7.061	7.149	2.480	5.809
	SSIM	0.747	0.790	0.753	0.895	0.824
	FID	29.070	25.835	49.602	18.630	15.592
Mask-Guided [2]	PSNR	19.661	23.647	21.376	29.000	21.800
	RMSE	7.294	6.736	7.136	2.454	5.884
	SSIM	0.752	0.819	0.776	0.902	0.815
	FID	18.431	27.199	16.364	18.112	18.871
MaskGAN [3]	PSNR	19.842	21.630	19.455	28.865	20.399 (collage [1])
	RMSE	7.982	7.485	7.621	2.524	9.011 (collage [1])
	SSIM	0.702	0.755	0.736	0.862	0.545 (collage [1])
	FID	32.568	26.387	27.254	19.142	16.256 (collage [1])

Table 6. **Quantitative Comparison for Fig. 7.** PSNR, RSME, SSIM, and FID scores on copy/paste (CP) and our generated (O) input pairs. The last column (from Fig. 8) lists component transfer results, with the top cell as our results.

G. Blending Comparison

As copy/paste operation is usually followed by blending for collages, we generated a blended image from copy/paste segments in Fig. 5 of the main text, and fed it as input to the same set of compared papers (Fig. 2). Results are still of lesser quality than using our generated masks and images as input, supporting our claims in Sec. 5.2.

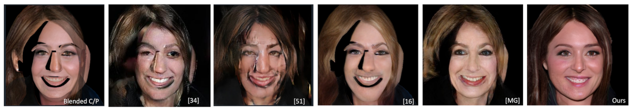

Figure 2. Results with **blended mask & image** by [1–3, 5, 6] (in order) versus ours (right).

H. Symmetry Coupling

Following our discussion from Sec. 5.3, Fig. 3 shows results without symmetry coupling. Although they look realistic at a first glance, different eye colors, gaze directions, and eyebrow styles give away their synthetic nature. Similarly for buildings, we coupled windows, cornices, and sills together for preserving patterns in random compositions. When those regions are selected randomly without following the same pattern, less dominant classes such as cornices and sills start to appear as phantoms on the buildings, as shown in the zoom ins.

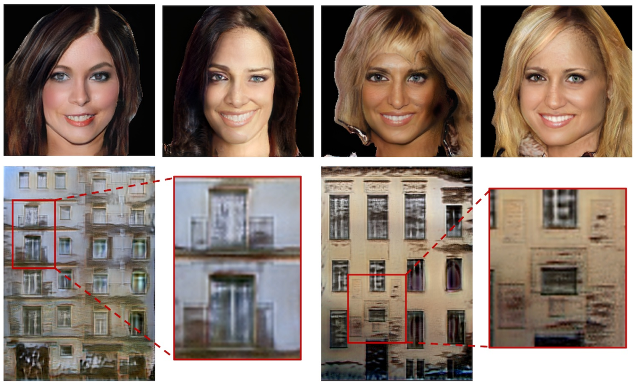

Figure 3. **Without symmetry coupling,** random faces have unmatched colors/brows (top) and buildings have phantoms (bot).

I. Extreme Cross-Dataset Reconstruction

In addition to the aforementioned reconstruction scores, here we add several interesting reconstructions where the source image is filtered, low resolution, or the actor has eyes closed in Fig. 5 (left). Moreover, we show hard cases with extreme color, scale, and pose variations in Fig. 5 (right), where the model is not trained on this type of data. In most cases, results look artistic rather than realistic, but not noticeably in the uncanny valley. Note that, the train and test datasets of Fig. 5 (right) are different as a stress case.

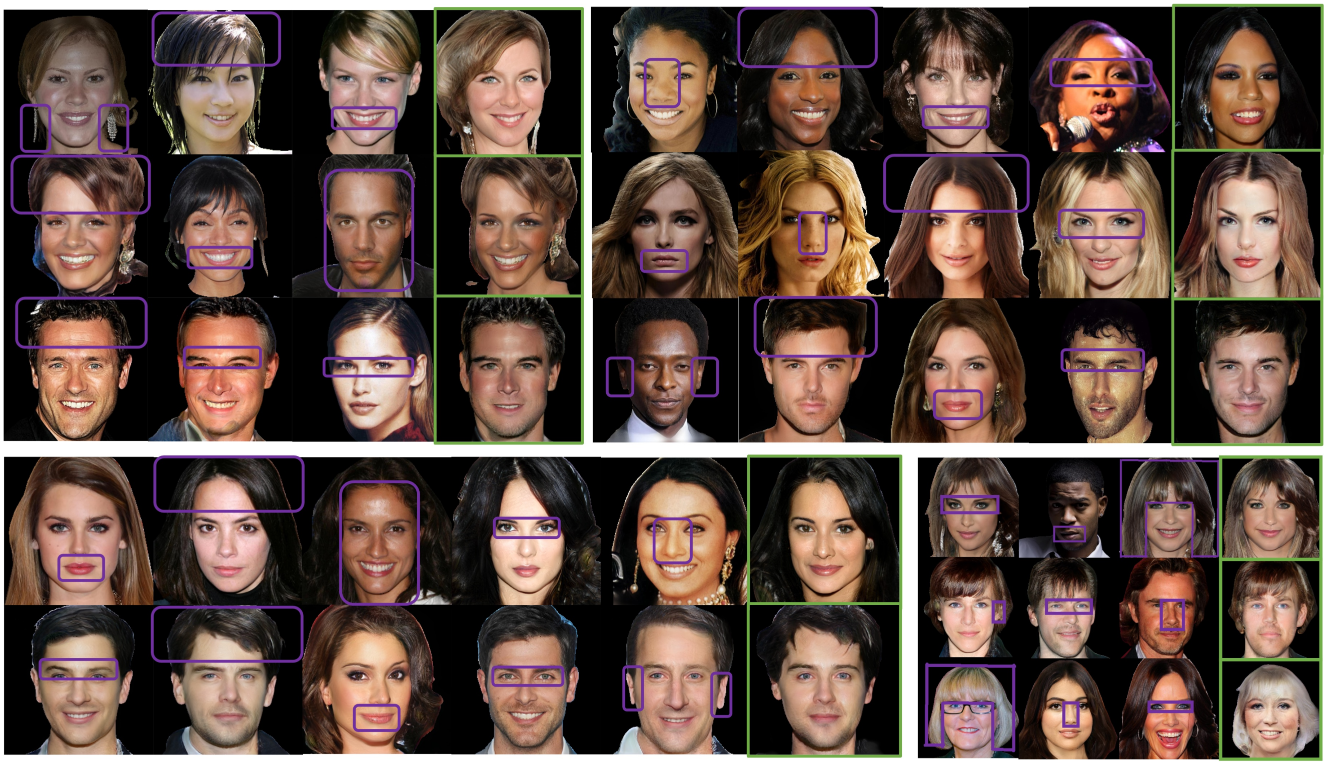

Figure 4. **Additional Results.** Purple-highlighted segments in the first 3, 4, or 5 images (per row) are used to synthesize new images (green). MixSyn combines source images with different skin tones, genders, illuminations, and regions, into a coherent realistic image.

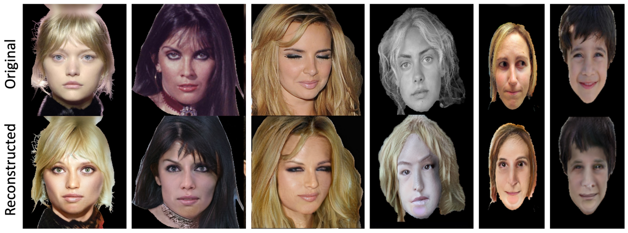

Figure 5. **Hard reconstructions** from CelebAMask-HQ (left half) and cross-dataset Helen (right).

J. Limitations

Some random combinations cause edge cases naturally. For example, if the face region is from an image with hair and the hair region is from a bald person, there may be a mismatch due to our network not seeing many similar samples (Fig. 6, first row). Similarly, when segments exhibit an extreme illumination/resolution/etc. change, such compositions create artistic effects (second row). Lastly, if there is a missing hat with hair, or if hat type is ignored, undesired images are generated (third row). We attest that user guidance hinders generating such random combinations with an interactive editing system.

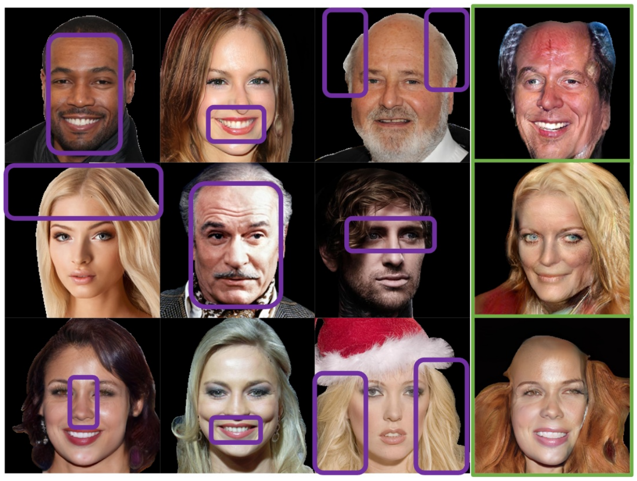

Figure 6. **Edge Cases.** Hair, illumination, and type mismatches.

K. Additional Face Synthesis Results

Fig. 4 demonstrates additional faces synthesized with MixSyn using varying number of source regions.

References

[1] Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in gans. In *International Conference on Learning Representations*, 2021. 5

[2] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[3] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4, 5

[4] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5654–5663, 2019. 4

[5] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[6] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5