Latents2Segments: Disentangling the Latent Space of Generative Models for Semantic Segmentation of Face Images

Snehal Singh Tomar A.N. Rajagopalan
Indian Institute of Technology Madras
snehal@smail.iitm.ac.in, raju@ee.iitm.ac.in

Abstract

With the advent of an increasing number of Augmented and Virtual Reality applications that aim to perform meaningful and controlled style edits on images of human faces, the impetus for the task of parsing face images to produce accurate and fine-grained semantic segmentation maps is more than ever before. Few State of the Art (SOTA) methods which solve this problem, do so by incorporating priors with respect to facial structure or other face attributes such as expression and pose in their deep classifier architecture. Our endeavour in this work is to do away with the priors and complex pre-processing operations required by SOTA multi-class face segmentation models by reframing this operation as a downstream task post infusion of disentanglement with respect to facial semantic regions of interest (ROIs) in the latent space of a Generative Autoencoder model. We present results for our model's performance on the CelebAMask-HQ and HELEN datasets. The encoded latent space of our model achieves significantly higher disentanglement with respect to semantic ROIs than that of other SOTA works. Moreover, it achieves a 13% faster inference rate and comparable accuracy with respect to the publicly available SOTA for the downstream task of semantic segmentation of face images.

1. Introduction

Multi-class semantic segmentation of facial regions of interest is central to various AR and VR applications. Yet, there exists a paucity of publicly available pre-trained models which can perform this task with reasonable accuracy. The promise of deep learning for general semantic segmentation has been explored by several works ([18], [3], [19], [16], [12], [1]) across a variety of scene settings. Face segmentation is a particularly challenging problem because of the irregular shapes, sizes, and textures of facial regions of interest. A category of literature ([23], [10]) solves the problem of segmenting out one region of interest (hair in most cases) by incorporating priors unique to that region in their architecture and losses.

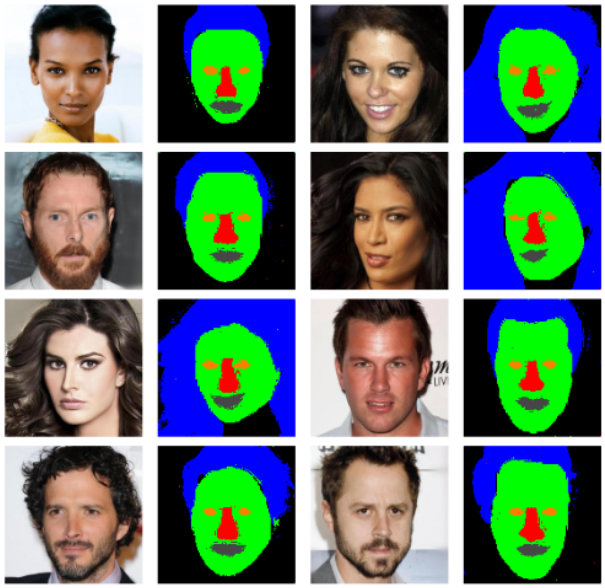


Figure 1. Representative semantic segmentation results on test images sampled from the CelebAMask-HQ dataset. Columns 1, 3 contain input face images and Columns 2, 4 contain their corresponding output segmentation maps, respectively. The color coding used for semantic regions in the segmentation maps is given by; blue: *hair*, green: *skin*, red: *nose*, orange: *eyes*, and grey: *lips + mouth*.

The SOTA when it comes to segmenting multiple regions of interest in face images are [11] and [22]. While, [11] relies on heavily pre-processed images, [22] incorporates relationships with facial expressions by learning graph representations. The computational overhead that pre-processing operations (warping) incur and the narrow scope of generalization of representation learning are key limitations of these works. These bottlenecks warrant the need for a pre-processing and prior independent approach that can generate fine-grained multi-class segmentation maps with a single forward pass over the input images. To this end, we propose the use of Generative Autoencoders

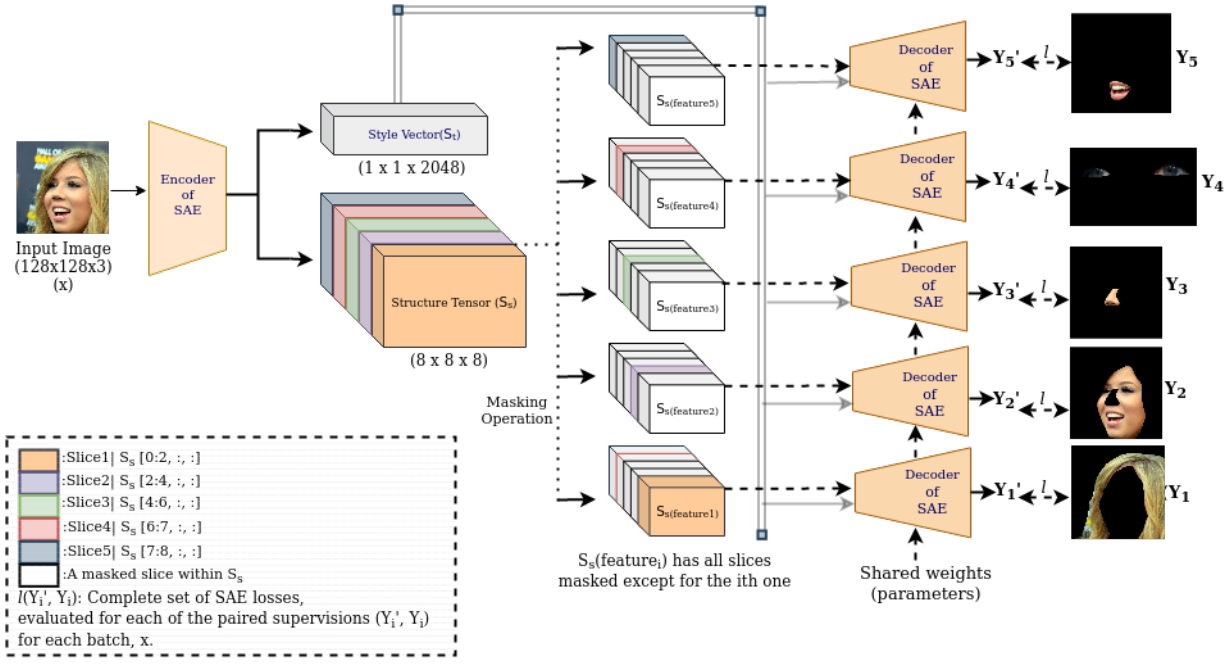


Figure 2. A schematic representation of our model's architecture and training pipeline. The schema used for slicing S_s and for formation of masked $S_{s_{feature_i}}$ have been annotated in the legend. l refers to the operation defined by Eq. 1.

(GAs) capable of producing regions of interest in a selective fashion, given an input image. GA models have shown promising results for tasks such as high fidelity reconstruction of images, style transfer, and style manipulation. However, the latent space of these models is heavily entangled and very high dimensional in nature. This limits their ability to enable spatially localised manipulation of images as a consequence of specific perturbations to their encoded latent space. The Swapping Autoencoder (SAE) [14] is an especially insightful work as it learns a latent representation with a neat distinction between structure and style information of the input image. Inspired by [14], we build upon its capabilities further in this work. Previously, generative models have been used extensively ([4], [24], [13], [20], [21]) for semantic image synthesis, which is the task of generating photo-realistic images that match the structure of a semantic segmentation map given as input. However, the problem at hand has not seen significant attempts in the past. The contributions of this paper can be summarized as below:

- We infuse a strong disentanglement with respect to the structure of semantic ROIs in re-generated images, in the latent space of SAE [14]. Ours is the first work which does so for any Generative Autoencoder model. We provide quantitative metrics for the extent of disentanglement achieved.
- We harness the disentangled nature of our model's latent space to perform the challenging downstream task of semantic face segmentation. Results obtained for this task are close to the current SOTA.
- Generating the segmentation map for a given ROI from an image belonging to a particular distribution using our model trained on a similar distribution amounts to a simple forward pass with appropriate masking (retention of a single non-zero slice corresponding to the ROI) applied to the latent space. This eliminates the need of any priors for semantic segmentation and underscores the applicability of our model to any generic semantic segmentation problem.

2. Methodology

The SAE [14] is a generative Autoencoder model which embeds the *structure* and *style* information present in input images ($H \times H \times 3$) into a *structure tensor* (S_s, having dimensions $H/16 \times H/16 \times 8$) and *texture vector* ($S_t$, having dimensions $1 \times 1 \times 2048$). The latent space $S = \{S_s, S_t\}$ serves as the point of initiation for our work. Our objective in this work, is to disentangle the tensor slices within S_s such that they correspond to the structure information of individual regions of interest, namely: hair, skin, nose, eyes, and (lips + mouth) in the reconstructed image. This, in effect, entails that *masking (setting to zero) all slices of S_s except one should should produce an image containing only its corresponding semantic region of interest, when decoded together with S_t*. This observation is key to our work and the results and experiments that follow, are based on it.

2.1. Network Architecture and Losses

Figure 2 depicts our model which is a parallelized version of the SAE [14] wherein, disentanglement with respect to semantic regions of interest is infused in the latent space. We achieve this by slicing the structure tensor and enforcing faithful reconstruction of the ROI by the decoder, when given the texture vector and a masked structure tensor (having only one nonzero slice that should correspond to the chosen ROI) as input. The correspondences that we have sought to develop are; Slice 1 (Channels 1 and 2 of S_s): ***Hair*** (R_1), Slice 2 (Channels 3 and 4 of S_s): ***Skin*** (R_2), Slice 3 (Channels 5 and 6 of S_s): ***Nose*** (R_3), Slice 4 (Channel 7 of S_s): ***Eyes*** (R_4), and Slice 5 (Channel 8 of S_s): ***Lips + Mouth*** (R_5). We did not optimize the chosen slicing scheme on the basis of amount of disentanglement obtained. Thus, our results are independent of the chosen slicing scheme. We treat each Siamese decoder together with the encoder as a separate SAE (including all its components viz. the encoder, generator (decoder in our case), discriminator, and patch co-occurrence discriminator) while computing losses. The loss for each correspondence pair is defined as:

$$\begin{aligned} l(Y_i', Y_i) = L_{\text{rec}}(Y_i', Y_i) + 0.5L_{\text{GAN, rec}}(Y_i', Y_i) \\ +0.5L_{\text{GAN, swap}}(Y_i', Y_i) + 0.5L_{\text{CooccurGAN}}(Y_i', Y_i) \end{aligned} \tag{1}$$

Here, Y_i' referes to the reconstruction obtained from the decoder for the latent space $\{S_{s_{feature_i}}, S_t\}$, Y_i refers to the ground-truth image containing R_i only, L_{rec} refers to the reconstruction(L1) loss, $L_{\text{GAN, rec}}$ refers to the non-saturating GAN loss, $L_{\text{GAN, swap}}$ refers to the non-saturating GAN loss for images generated post swapping, and $L_{\text{CooccurGAN}}$ refers to the patch co-occurrence discriminator loss as defined in [14]. For every batch of training data, parameters of the encoder and decoder (all Siamese decoders share the same parameters) are optimized using the following overall loss:

$$L_{\text{overall}} = \sum_{i=1}^{5} 0.2 \cdot l(Y_i', Y_i) \tag{2}$$

Training was initiated using the pre-trained weights provided by [14], post training on the FFHQ dataset [6]. Inferring segmentation maps from our model amounts to a simple forward pass over the network with appropriate masking applied to S_s in the latent space, depending upon the ROI for which the segmentation has to be performed.

2.2. Data Preparation

We have used images (resized to $128 \times 128 \times 3$ dimensions) from the CelebAMask-HQ [9] and HELEN [8] (contains images in the wild) datasets for training and evaluation. A batch of training data comprised of $\{X, Y_1, ..., Y_5\}$ where X denotes a batch of input images and Y_i denotes a batch of region specific images, R_i. We sampled only those images from the CelebAMask-HQ dataset which had annotations for all the intended ROIs and split them into 27016 training and 862 test images. The HELEN dataset was used as made available.

3. Experiments

Since the objective of this work is to disentangle the latent space of a GA model so as to develop strong correspondences between latent slices and semantic regions of interest in order to harness the same for the downstream application of semantic segmentation of face images, we analyse our model's performance primarily with respect to two criteria, namely the *Degree of Disentanglement* achieved and the *accuracy* of predictions for the chosen downstream task.

3.1. Degree of Disentanglement

Degree of Disentanglement implies the extent to which changes made within a chosen slice of the latent representation (S_s) affect the pixel intensities in the desired semantic region of interest (R_i), without causing any deviation to those in other ROIs. We propose the use of an activeness measure inspired by [15] to quantify the degree to which a latent space conforms to this desired attribute. We define the activeness of a latent tensor masked in accordance with the i^{th} slice of S_s and denoted as $S_{s_{feature_i}}$, with respect to a particular semantic ROI (R_j) as:

$$A_{ij} = \mathbb{E}_{n,x}(\sigma^2((Decoder(S_{s_{feature_i}}, S_t) \cdot Mask_{R_j}))) \tag{3}$$

where,

$$Encoder(x) = \{S_s, S_t\} \tag{4}$$

$$S_s \cdot Mask_{R_j} = S_{s_{feature_j}} \tag{5}$$

In essence, activeness of $S_{s_{feature_i}}$ with respect to R_j for a given input x, is the expectation of variance observed in R_j due to addition of noise ($n \in 0.01 \cdot N(0, 1)$) to $S_{s_{feature_i}}$, taken over all the added noise tensors. We also define the average activeness map (Map) for a data distribution, such that $Map_{ij} = \mathbb{E}_x(A_{ij})$.

Figure 3 depicts the average activeness map for our model obtained on test data sampled from the CelebAMask-HQ dataset. Since the map (matrix) is nearly diagonal, $S_{s_{feature_i}}$ controls the pixel intensity levels in R_j significantly, only if $j = i$. Thus, the claim made in section 2.1 regarding the correspondences that our model develops, stands validated. We define the ratio of sum of diagonal elements of the average activeness map to that of sum of all elements of the average activeness map as the ***Activeness Compaction Score*** **(ACS)**. The ACS for our model with regard to the CelebAMask-HQ dataset [9] was found to be **0.8186** which suggests a nearly diagonal nature of the obtained Map. Thus, it is evident that the slices of our model's

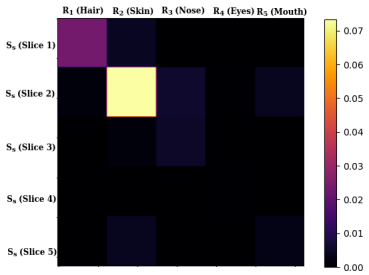


Figure 3. Our model's average activeness map obtained on test data sampled from the CelebAMask-HQ dataset. The near diagonal nature of the map suggests a high level of disentanglement among the slices of S_s with respect to the intended semantic ROIs.

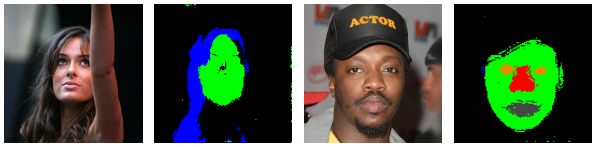
Method	Latent Dimensions	Slicing Scheme	ACS↑
Pixel2Style2Pixel [17]	18×512 [18, 512]	[0:4, :] : Hair [4:8, :] : Skin [8:12, :] : Nose [12:16, :] : Eyes [16:18, :] : Lips + Mouth	0.189
StyleGAN2-ADA [5]	18×512 [18, 512]	[0:4, :] : Hair [4:8, :] : Skin [8:12, :] : Nose [12:16, :] : Eyes [16:18, :] : Lips + Mouth	0.194
Swapping Autoencoder [14]	$8 \times 8 \times 8$ [8, 8, 8]	[0:2, :, :] : Hair [2:4, :, :] : Skin [4:6, :, :] : Nose [6:7, :, :] : Eyes [7:8, :, :] : Lips + Mouth	0.259
Latents2Segments (Ours)	$8 \times 8 \times 8$ [8, 8, 8]	[0:2, :, :] : Hair [2:4, :, :] : Skin [4:6, :, :] : Nose [6:7, :, :] : Eyes [7:8, :, :] : Lips + Mouth	**0.819**

Table 1. Comparative analysis of our model's latent space with that of SOTA Generative Models which either encode or project input images onto a structured latent space on the basis of ACS using test images from CelebAMask-HQ [9] dataset. Our model's latent space is the most disentangled and by a large margin, with respect to semantic ROIs.

Method	F1 Score ↑					Computation Time(ms) ↓
	Hair	Skin	Nose	Eyes	Skin+Mouth	
Modified BiSeNet	0.9524	0.89	0.931	0.81	0.741	142.69
Ours	0.7803	0.8532	0.7605	0.578	0.6715	**124.00**

Table 2. F1 scores and time taken per input image for segmentation with respect to different semantic ROIs obtained on test images from the CelebAMask-HQ dataset. Our model's disentangled latent space yields performance comparable to publicly available pre-trained SOTA (Modified BiSeNet [2]) for most ROIs and the rate of inference is faster.

latent space have a direct correspondence with pixel intensities in the intended semantic regions of interest only. We present a comparative analysis of our model's latent space with that of several SOTA works in Table 1. The slicing schema chosen for these experiments were taken to be close to ours in order to maintain consistency. Our model's latent space has the most disentanglement (with respect to semantic ROIs) infused in it.



(a) (b)

Figure 4. Challenging instances from test images belonging to (a) HELEN (in the wild) and (b) CelebAMask-HQ dataset. Our model performed well despite the occlusions (regions not belonging to any semantic ROI), being present. The color coding used for the predicted segmentation maps is the same as that used for Figure 1.

3.2. Segmentation Accuracy

Qualitative results have been presented in Figure 1 (refer to section 1) and Figure 4. We chose to disentangle two large slices (each containing 4 channels) of S_s, with respect to *Hair* and *Skin*, respectively, while working with the HELEN dataset as the number of training images was lesser than that required to infuse disentanglement within several slices. We compare the accuracy of our model's predicted semantic labels for different ROIs and its rate of inference with SOTA, on the basis F1 scores and computation time per input image in Table 2. Our model is faster and comparable to SOTA for the chosen downstream task. From Table 2, we infer that certain entanglements in the latent space of our model are essential for near perfect re-generation of structure information by the decoder, since it inherits from the StyleGAN2 [7] architecture. Therefore, there is a trade-off between the extent to which S_s is disentangled, and the segmentation accuracy obtained. Since, this work focuses on disentanglement, we have optimized only the amount of disentanglement achieved, and will take up the characterization of this trade-off as a future work. We also observe that the accuracy of predicted segmentation maps and the number of classes for which segmentation is feasible, has a direct correlation with the amount of varied and well-annotated training data available.

4. Conclusion

In this work, we proposed and evaluated a method to disentangle the latent space of Generative Autoencoder models with respect to semantic ROIs. We illustrated its applicability to the downstream task of semantic segmentation of face images. Our model outperforms SOTA in terms of disentanglement and is faster while being comparably accurate in performing the downstream task. Our model shall find tremendous utility, especially in AR/VR applications that require selective control over semantic ROIs as a prerequisite since it is entirely prior-agnostic.

Acknowledgement: Support from Institute of Eminence (IoE) project No. SB20210832EEMHRD005001 is gratefully acknowledged.

References

[1] Mehmet Aygun, Aljosa Osep, Mark Weber, Maxim Maximov, Cyrill Stachniss, Jens Behley, and Laura Leal-Taixe. 4d panoptic lidar segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5527–5537, June 2021. 1

[2] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9711–9720, 2021. 4

[3] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[5] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 4

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. 3

[7] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[8] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang. Interactive facial feature localization. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 679–692, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 3

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

[10] Alex Levinshtein, Cheng Chang, Edmund Phung, Irina Kezele, Wenzhangzhi Guo, and Parham Aarabi. Real-time deep hair matting on mobile devices. *CoRR*, abs/1712.07168, 2017. 1

[11] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[12] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L. Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[13] 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)*, June 2019. 2

[14] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Advances in Neural Information Processing Systems*, 2020. 2, 3, 4

[15] William Peebles, John Peebles, Jun-Yan Zhu, Alexei A. Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 3

[16] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 1

[17] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 4

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 1

[19] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 1

[20] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. Diverse semantic image synthesis via probability distribution modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7962–7971, June 2021. 2

[21] Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, Gang Hua, and Nenghai Yu. Semantic image synthesis via efficient class-adaptive normalization. *CoRR*, abs/2012.04644, 2020. 2

[22] Gusi Te, Yinglu Liu, Wei Hu, Hailin Shi, and Tao Mei. Edge-aware graph representation learning and reasoning for face parsing. In *European Conference on Computer Vision*, pages 258–274. Springer, 2020. 1

[23] Andrei Tkachenka, Gregory Karpiak, Andrey Vakunov, Yury Kartynnik, Artsiom Ablavatski, Valentin Bazarevsky, and Siargey Pisarchyk. Real-time hair segmentation and recoloring on mobile gpus. *CoRR*, abs/1907.06740, 2019. 1

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2