Supplementary material:Decoupling identity and visual quality for image and video anonymization

Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé

Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany

Abstract. In this supplementary material, we provide the implementation details of our method *(Section 1)*, detail the datasets *(Section 2)*, provide more qualitative results in higher resolution *(Section 3)*, qualitatively compare the results of our method with those of CIAGAN *(Section 4)*, show more qualitative ablation studies *(Section 5)*, demonstrate anonymization on temporal data *(Section 6)*, show additional results on full-body anonymization *(Section 7)*, and provide the network's architecture *(Section 9)*.

1 Implementation details

We train our network on 128×128 resolution images for 60 epochs using the Adam optimizer [3] with $1e{-}5$ learning rate. We set the beta hyperparameters β_1, β_2 to 0.5 and 0.9. We train the network for two days in a single GPU (TitanX).

We train our HarmonizationNet on a proxy task using the following augmentations: random resized crop, random changes on the hue and brightness, random color balance, and random gaussian noise.

Computation time. In Tab. 1, we show the generation time (in seconds) of our method and compare it with CIAGAN. For completeness, we also show the quality of the generation (Tab. 4 in the main paper). We show that in low-resolution generation (x128), CIAGAN is significantly faster than our method (0.01927 seconds vs 0.03784 seconds). However, in higher resolutions, the difference gets lower, and in x512 super-resolution the difference is marginal. Considering the massive difference in quality, and the relatively fast time of our method (23 frames per second), we conclude that our method significantly improves the results by needing a slightly longer generation time (0.003 more seconds in super-resolution), thus showing an good quality-time tradeoff.

Method	FID(↓)			Sec. (↓)		
	x128	256	x512	x128	256	x512
CIAGAN + SR method	35.89	38.82	37.98	**0.019**	**0.030**	**0.041**
Ours	**10.49**	**8.41**	**11.60**	0.038	0.039	0.044

Table 1. Comparison of our method with CIAGAN in both quality and time performance. Lower (↓) results for FID imply a higher generation quality, and (↓) results for Sec. imply a faster generation.

2 Datasets

Datasets. We perform training and experiments on 6 public datasets:

- **CelebA** [5] The dataset consists of $202,599$ face images of $10,177$ unique identities. We use the aligned version where each image is centered on a point in-between person's eyes, and then padded and resized to have 178×218 resolution, while maintaining original face proportions. We sampled identities that contain at least 20 images.
- **CelebA-MaskHQ** [5] The dataset consists of $30,000$ face images with corresponding 19 classes of face part segmentations. We train face segmentation model on this dataset.
- **Labeled Faces in the Wild (LFW)** [2] The dataset consists of $6,000$ pair images, split in 10 different splits, where half of the pairs contain images of the same identity, and the remaining pairs consist of images that have different identities.
- **FaceForensics++** [7] dataset has 1000 videos. We use 200 of them for testing to evaluate temporal consistency and 800 for fine-tuning our model.
- **AFLW2000** [4] is a challenging dataset consisting of 2000 images with face rotation annotations. It contains a significant number of faces with extreme poses.
- **MOTS** [8] Our method can also be adapted to work in other domains such as full-body anonymization. We use whole a silhouette mask and body joints as the representation. We sample 113 video sequence, each containing a different person.

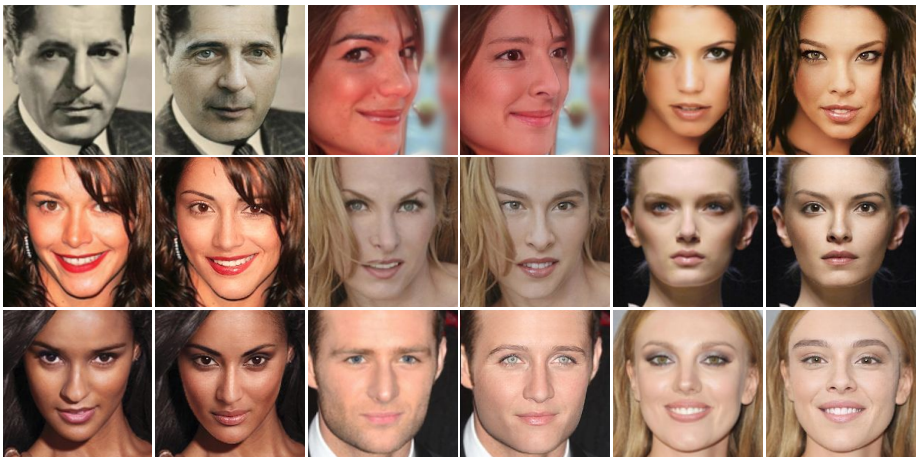

Fig. 1. Qualitative results in x512 resolution. In each pair, the first image is the original image given in x128 resolution, while the second image is the image generated by our method in higher resolution.

Fig. 2. Qualitative diversity results in x512 resolution. In each pair columns, the first column is the output of the AnonymizationNet, while the second column is the final blended output of the HarmonizationNet.

3 Higher resolution

We provide further qualitative results of our method on x512 resolution in Fig. 1. Our generated results still do not have the exact same skin tone as the input images and might contain small visual artifacts. However, the generated images preserve the pose, are sharp with realistic facial details, and look significantly different compared to the original input image. Additionally, we provide diverse results of out method on x512 resolution in Fig. 2, where we show the results of both AnonymizationNet and HarmonizationNet.

4 Comparison

In Fig. 3 we show more qualitative results of our method and compare them with the results of CIAGAN [6]. As can be seen, CIAGAN [6] has problems with small occlusions and extreme poses, problems that are mitigated by our method.

5 Ablation studies

In Fig. 4, we use different type of inputs to the HarmonizationNet. We see that our blending network is invariant to color changes and can even process inputs with a clear domain gap from the training data. Intuitively, the network is trying to filter out the general shape from the input and then fill it with a generated texture.

Additionally, in Fig. 5 and Fig. 6 we show the qualitative results of our designed model and compare it with two other model configurations in CelebA

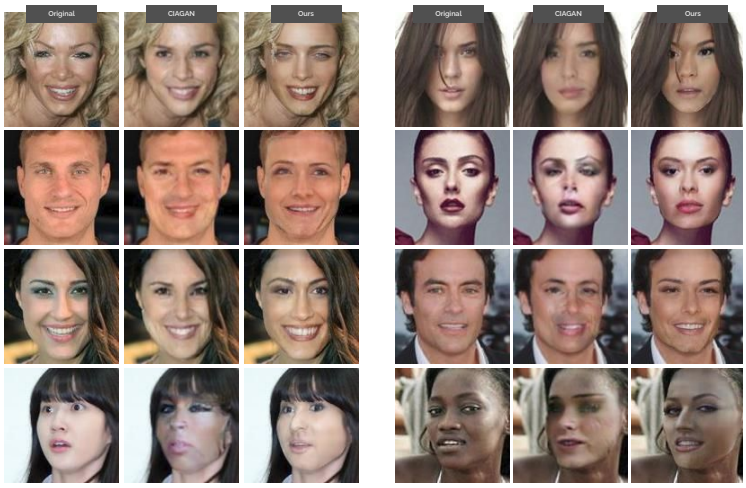

Fig. 3. Qualitative results on CelebA. In each triplet, the first image is the original image, the second image is generated by CIAGAN and the third image is generated by our method.

and Celeba-HQ dataset [5]. We also provide qualitative on the diversity in Fig. 7 and Fig. 8. As seen from all figures, the model without decoupling exhibits noticeable artifacts and noise in the output images. We see that our two-step method reaches by far the best qualitative results, especially in Celeba-HQ where the other design choices do not perform well.

6 Temporal consistency

As mentioned in the main paper, we improve the temporal consistency by simply adding additional input to the HarmonizationNet and to its discriminator. The generator has two encoders that take the current frame and anonymized version of the previous frame (or an empty image, if it is the first frame) as inputs. We concatenate both encoder embeddings at the bottleneck and pass it to the decoder part. The discriminator takes a concatenation of output of the current frame and output of the previous frame. This temporal discriminator trains on identifying both a realistic generation for each frame and a realistic temporal difference between frames. Note that we can not use temporal consistency module on the segmentation network due to the lack of face segmentation ground truth on temporal data. Due to training only on image dataset, the segmentation output is temporally inconsistent which reflects on the final output.

We provide a video file where we explain our method, show image results on CelebA-Mask-HQ and temporal results on FaceForensic datasets [7]. The narration was done by a synthesized voice using off-the-shelf text-to-speech model. The first part of the video shows four different anonymizations in three video

Fig. 4. Qualitative results on different type of input to the HarmonizationNet. In the first column we show the original images given as an input to our framework. The first and third rows are differently processed outputs of the AnonymizationNet that we give as an input to the HarmonizationNet: 2nd column - as it is, 3rd column - random change in the hue, 4th column - processed by an edge detector, 5th column - processed by the laplacian edge detector. The second and fourth rows are final outputs of the HarmonizationNet.

sequences. Each sequence contains an original frame, a segmentation estimate, and anonymized output of our model. The second part shows a comparison with and without temporal consistency. Each sequence contains an original frame, a segmentation map estimate, anonymized outputs without and with the temporal module. As can be seen, the temporal module shows smoother output with less color jittering. Still, the consistency is not perfect, mainly due to small segmentation differences between different frames - lack of temporal awareness for segmentation network. The last part of the video is a gallery of several video sequences. Each sequence contains an original frame and our anonymized version in pairs.

7 Different domain

We train our method on MOTS dataset [8] with 70 video sequences. We use silhouette masks and estimate body joints, using OpenPose [1], and give them as input to AnonymizationNet. We show the qualitative results of our method on full-body anonymization in Fig. 9. Our method maintains the same posture as the original while generating a new appearance, with HarmonizationNet acting like a refining network.

Fig. 5. Qualitative ablation results on CelebA [5] dataset where we compare our chosen model with the three other models.

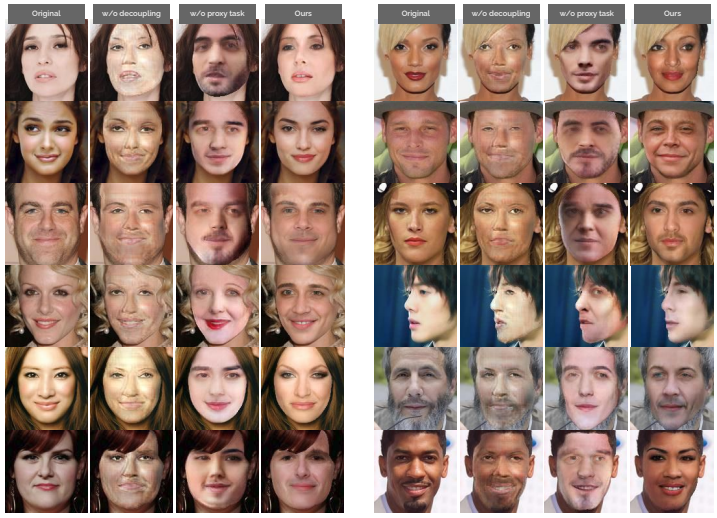

Fig. 6. Qualitative ablation results on CelebA-HQ [5] dataset where we compare our chosen model with the three other models.

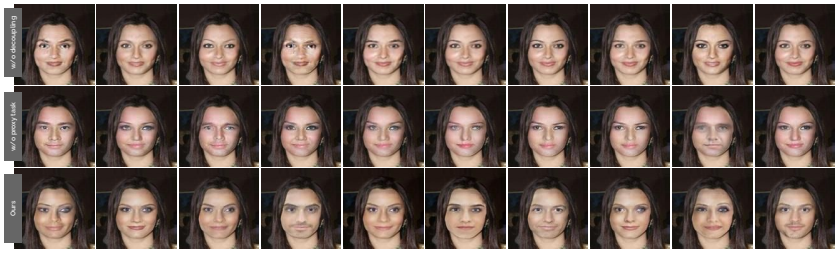

Fig. 7. Qualitative diversity results on CelebA [5] dataset where we compare our chosen model with the three other models.

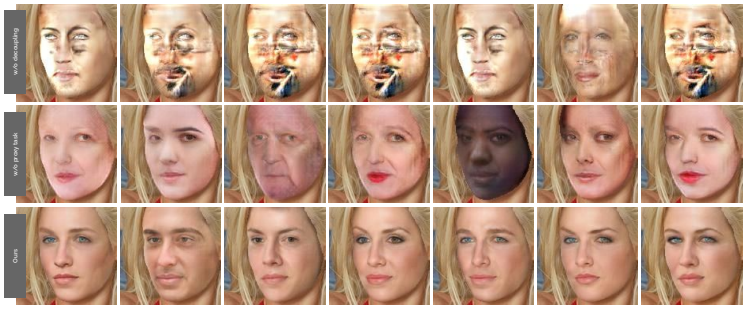

Fig. 8. Qualitative diversity results on CelebA-HQ [5] dataset where we compare our chosen model with the three other models.

8 Limitation

The main limitation of our method is its reliance on good segmentation masks. If the segmentation network is unable to segment properly the original face, then the quality of the result will be lower. Similarly to the previous anonymization methods, very extreme poses still might cause generation artifacts. Furthermore, we also observed that our method works worse in images taken in a low light scenario. Both limitation are mainly due to the lack of extreme poses and challenging scenes in the training data. Such photographic biases are quite notable in the celebrity focused datasets that we mainly use.

9 Network's architecture

In order to make the results of the paper reproducible, in Figures 10 and 11 we give the detailed architecture for AnonymizationNet and HarmonizationNet.

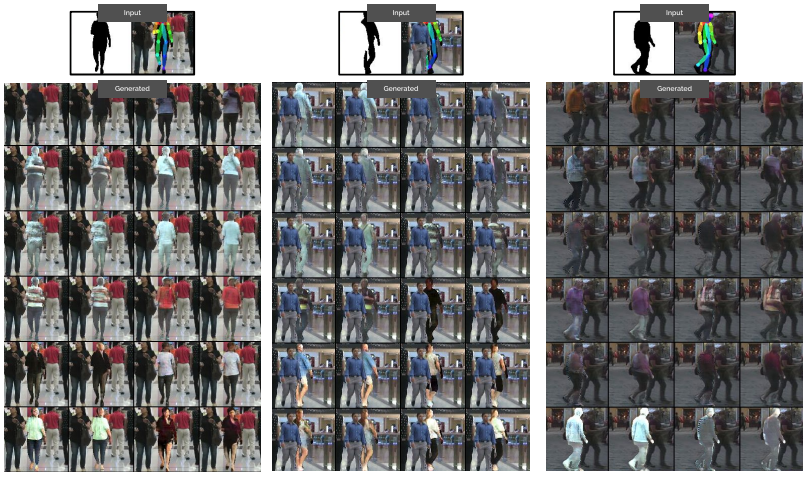

Fig. 9. Qualitative results in full-body anonymization using MOTS dataset [8].

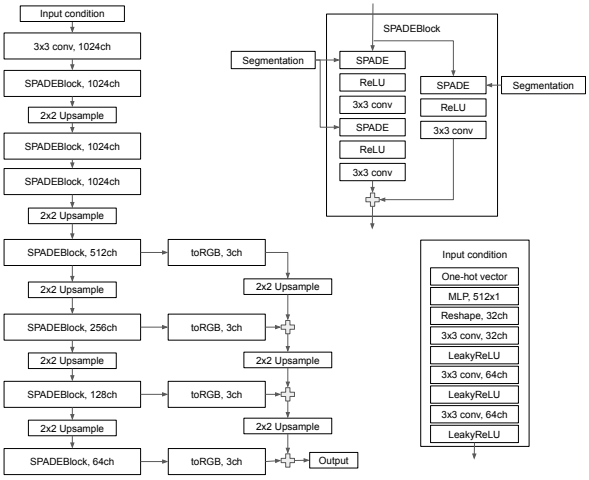

Fig. 10. The architecture of AnonymizationNet

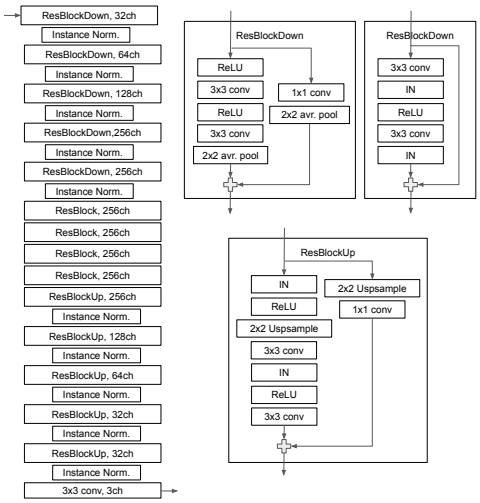

Fig. 11. The architecture of HarmonizationNet

References

1. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
2. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
3. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2014)
4. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: International Conference on Computer Vision Workshops (ICCVW). pp. 2144–2151 (2011)
5. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: International Conference on Computer Vision (ICCV) (2015)
6. Maximov, M., Elezi, I., Leal-Taixé, L.: CIAGAN: conditional identity anonymization generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5446–5455 (2020)
7. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: International Conference on Computer Vision (ICCV). pp. 1–11 (2019)
8. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: MOTS: multi-object tracking and segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7942–7951 (2019)