-Supplementary Material- FacialGAN: Style Transfer and Attribute Manipulation on Synthetic Faces

Ricard Durall[1,2]
ricard.durall.lopez@itwm.fraunhofer.de
Jireh Jam[3]
jireh.jam@stu.mmu.ac.uk
Dominik Strassel[1]
dominik.strassel@itwm.fraunhofer.de
Moi Hoon Yap[3]
m.yap@mmu.ac.uk
Janis Keuper[1,4]
janis.keuper@hs-offenburg.de

[1] Fraunhofer ITWM, Kaiserslautern, Germany
[2] IWR, University of Heidelberg, Heidelberg, Germany
[3] Manchester Metropolitan University, Manchester, United Kingdom
[4] IMLA, Offenburg University, Offenburg, Germany

1 Direct Comparison with StarGANv2 and MaskGAN

In this section, we present a detailed comparison between current state-of-the-art image-to-image facial translation models. In particular, we benchmark our FacialGAN approach against StarGANv2 [1] and MaskGAN [2]. Table 1 shows a concise summary, emphasizing the main properties of each model.

We start outlying the differences of our proposal regarding StarGANv2 [1]. For the preservation of the target attributes, to guarantee the identity, [1] employs a pretrained network based on an adaptive wing loss [4] that generates heatmaps. These heatmaps, however, do not provide fine-grained control over the attributes. They just signalize to the generator whether to keep or not all the facial attributes when applying the styles. Besides this limitation, the heatmaps require the input images to be vertically and horizontally aligned to have the eyes at the centre. Otherwise, attribute preservation may fail. Furthermore, the system relies on domain-specific modules to enforce semantic information, such as the output's gender. As a result, the architecture and its performance are dependent on the number of different semantic labels with which the system works, limiting its scalability. Overall, [1] can successfully synthesize images of various styles, but with scalability issues and no pixel-wise control. FacialGAN, on the other hand, can manipulate facial attributes at the pixel level while also applying styles from reference images. To accomplish this, we have made a number of changes to our model. First, we modify the generator to handle semantic mask labels as input. This modification has two direct consequences: (1) it removes the heatmaps dependency, greatly simplifying the system, and (2) it allows the pixel information to be used to control the attributes. To make use of such information, we add a segmentation

Method	scalability	style trans.	attr. manipulation	editing control	train stages
StarGANv2 [1]	limited	good	limited	none	one
MaskGAN [2]	good	limited	good	good	two
Ours	good	good	good	good	one

Table 1: Summary of main differences between our model, StarGANv2 and MaskGAN.

network, a modified U-Net [3], and a customized loss function that interacts with semantic input signals. More specifically, we propose a new local segmentation loss that propagates informative gradients only from the region of interest, i.e. the target pixel-wise attributes. In this way, we ensure that the generated output adheres to the mask specifications. Finally, the addition of geometry information (segmentation mask) might be seen as an alternative to semantic information, allowing us to scale up our model to work with more attributes without adding complexity, i.e. dedicated modules.

The differences between FacialGAN and MaskGAN [2] are also notable. Both models are radically different in terms of design. [2] trains in a fairly complex setup, which is divided into two stages with three different networks. In the first stage, the model learns the mapping between the semantic mask and the output image. Once it has converged, they start training the second stage, where the method learns to model the user editing behaviour when manipulating semantic masks. Additionally, there is an encoder-decoder architecture called MaskVAE, which is in charge of generating geometrical masks for training, that needs to be pretrained beforehand. All the training relies on the optimization of an adversarial, feature and perceptual loss. Despite synthesizing successfully images with different pixel-wise control, [2] is not designed for advanced style transfers. In other words, it lacks the ability to apply morphological changes when applying style, resulting in a very limited approach to style transfer. FacialGAN, on the other hand, is capable of extracting and applying cutting-edge style transfers, as well as modifying the geometry of the image if needed. We accomplish this by using a one-step training method in which the model learns the style while manipulating the semantic masks. As a result, FacialGAN has a more compact training that produces superior distribution-level metrics (FID and LPIPS). Identity preservation and attribute manipulation are two other areas where our proposal outperforms [2]. Mainly due to our new local segmentation loss, that forces the generator to focus only on the target regions, leaving the rest unmodified. Thus, the outcomes preserve unaltered the main features, respecting the identity, with pixel-wise control in case the user desires to change some attributes.

2 Facial Editing Toolbox

We propose an interactive facial toolbox that allows easy manipulation of both styles and attributes. The user chooses the source and reference images, and our model generates the desired combination in the desired direction. In addition, the user can change the default mask, and the changes are reflected in the output. We believe that such a tool can be very useful for validating results and allowing practitioners to continue to improve on facial manipulations. The source codes, pretrained models, and facial editing toolbox can be found on Github. We also provide a video tutorial where we show how to use the toolbox. It is available on Youtube.

3 Additional Results

Figure 1 displays additional comparisons with the baseline models on reference-guided generation. Furthermore, we provide additional image synthesis results where we apply an extensive variety of styles and mask's modifications. Figure 2, Figure 3 and Figure 4 display generated results when mouth, nose or eyes (eyebrows) have been altered through their segmentation masks. The results, where more than one attribute was changed, are shown in Figure 5. Finally, we run a gender translation experiment where the source and the reference images are the same. This way, if we exchange the original gender, we still get the "same" person with the same style, but with the opposite gender. Results are displayed in Figure 6.

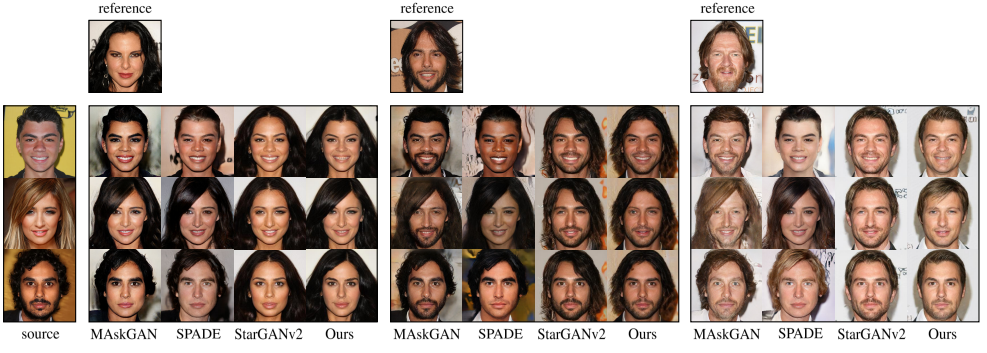

Figure 1: Qualitative comparison of style transfer on reference-guided generation.

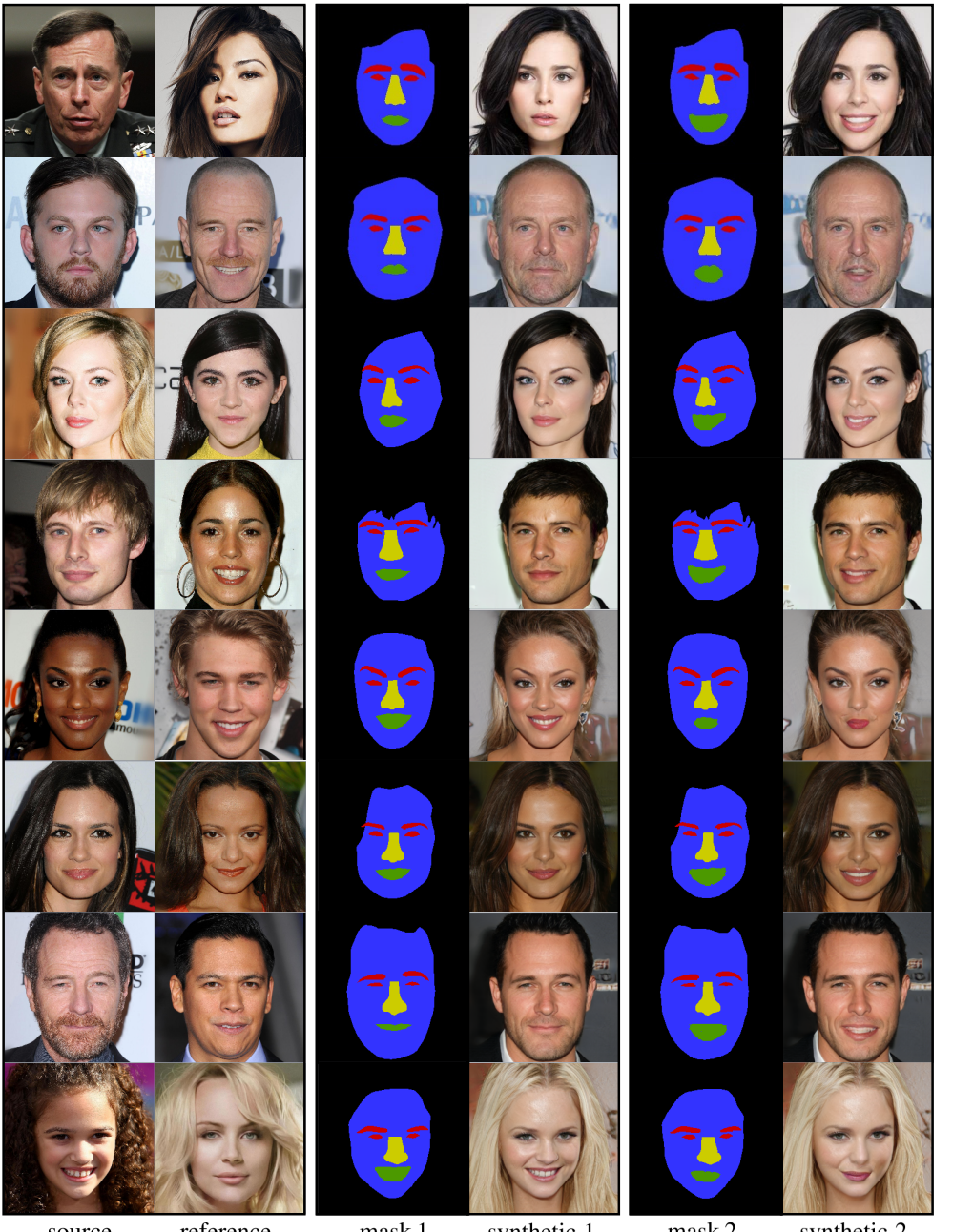

Figure 2: In this picture, we show that our model is able to learn to transform a source image to reflect the style of a given reference image while being consistent with the semantic mask. In particular, we only modify the mask of the *mouth*. The source and style reference images appear in the first two columns, whereas the respective transformation masks are given in column 3 and 5. The columns 4 and 6 show the generated images.

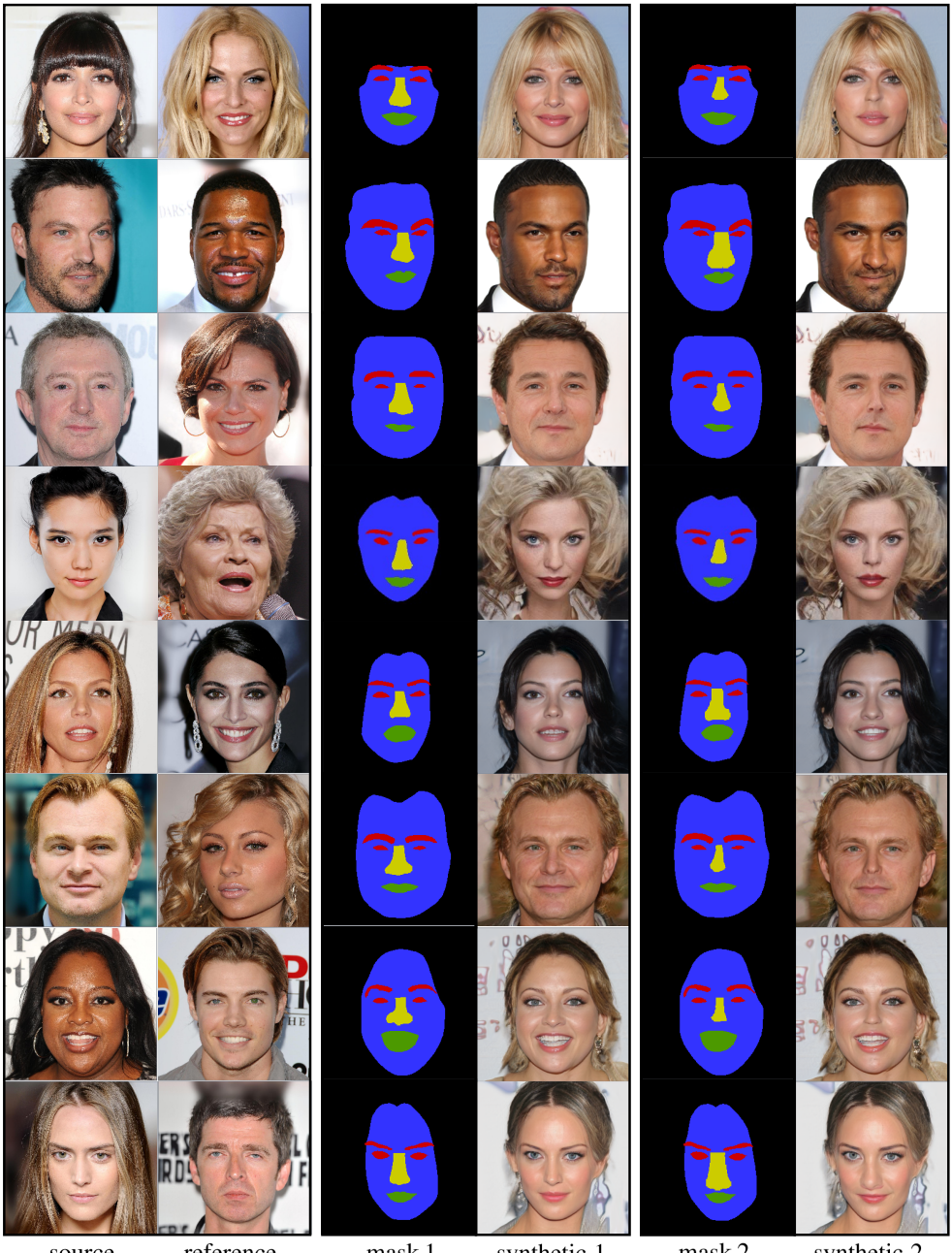

Figure 3: Our model is able to learn to transform a source image to reflect the style of a given reference image while being consistent with the semantic mask. In particular, we only modify the mask of the *nose*. The source and style reference images appear in the first two columns, whereas the respective transformation masks are given in column 3 and 5. The columns 4 and 6 show the generated images.

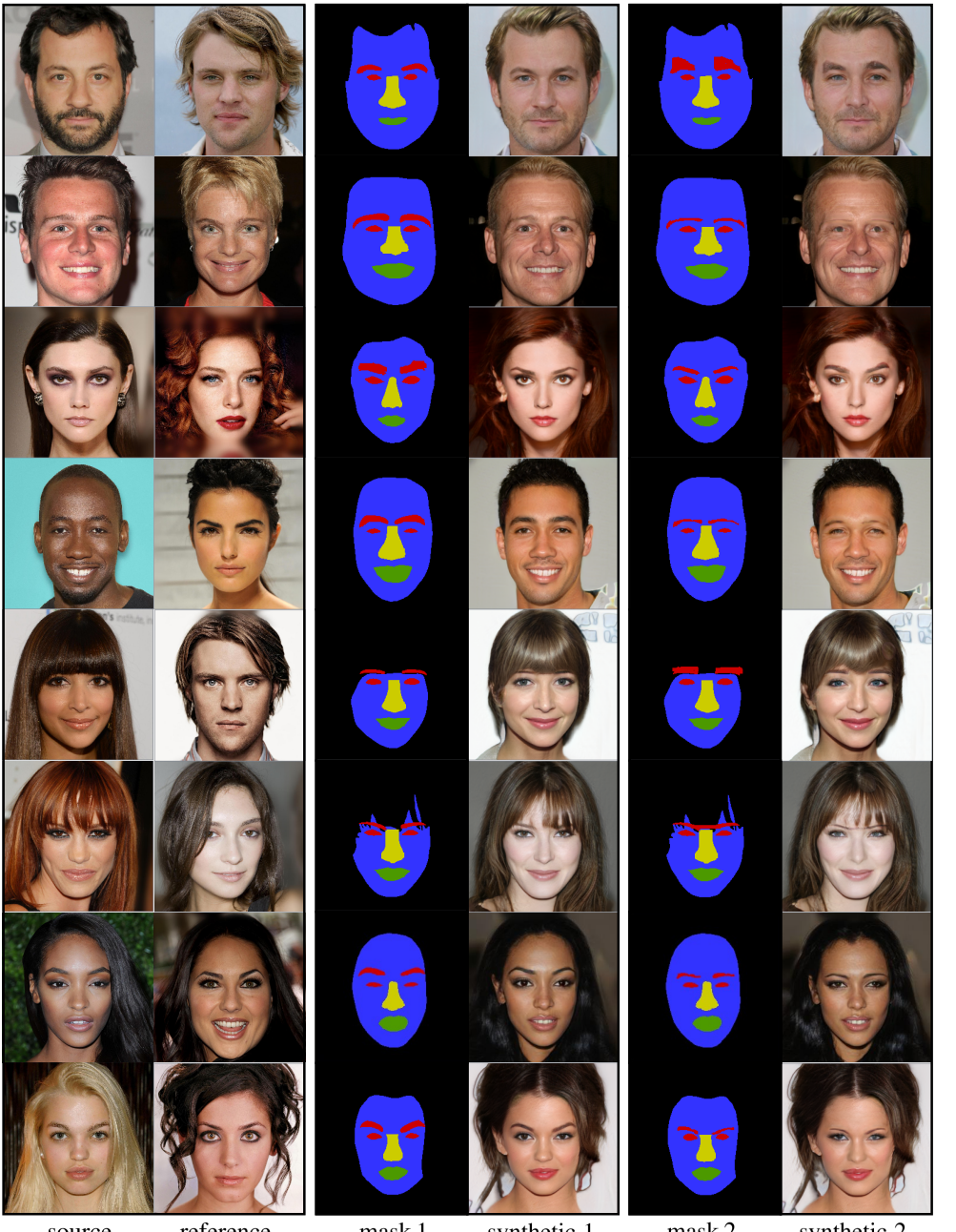

Figure 4: Our model is able to learn to transform a source image to reflect the style of a given reference image while being consistent with the semantic mask. In particular, we only modify the mask of the *eyes* (*eyebrows*). The source and style reference images appear in the first two columns, whereas the respective transformation masks are given in column 3 and 5. The columns 4 and 6 show the generated images.

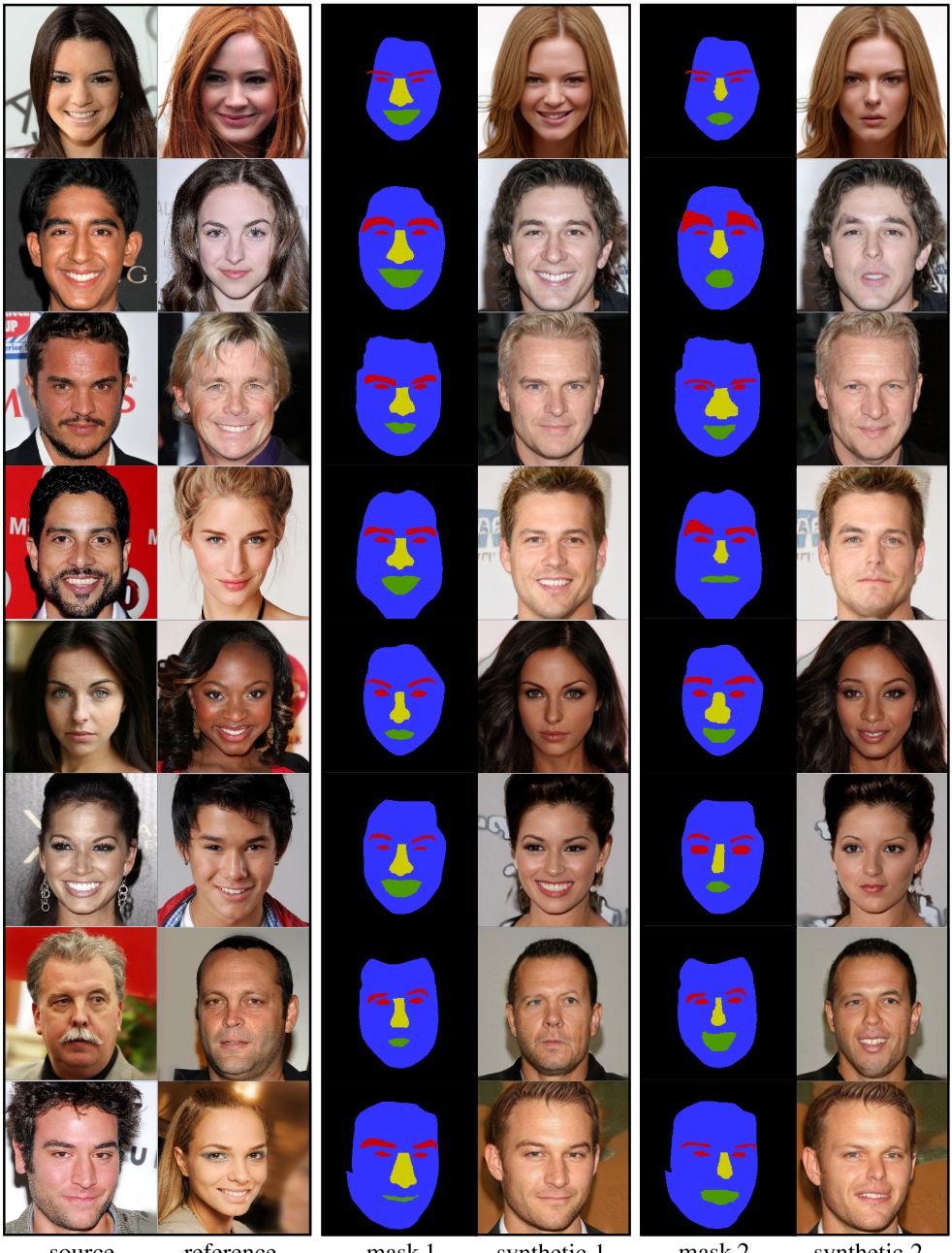

Figure 5: Our model is able to learn to transform a source image to reflect the style of a given reference image while being consistent with the semantic mask. In particular, we modify more than one attribute of the mask. The source and style reference images appear in the first two columns, whereas the respective transformation masks are given in column 3 and 5. The columns 4 and 6 show the generated images.

Figure 6: Our model is able to apply gender translation. The source and style reference images are the same, and appear in the first columns of each block. The respective synthetic results from the gender transformation are given in second columns.

References

[1] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[2] 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*, pages 5549–5558, 2020.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[4] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6971–6981, 2019.