# GAN-Control: Explicitly Controllable GANs —Supplementary—

Alon Shoshan Nadav Bhonker Igor Kviatkovsky Gérard Medioni
Amazon
{alonshos, nadavb, kviat, medioni}@amazon.com

## 1. Introduction

In the supplementary material we present experiments on images of dog (Sec. 2), further implementation details (Sec. 3), method analyses (Sec. 4), qualitative photorealism comparison to other methods (Sec. 5), additional ablation study details (Sec. 6), limitations (Sec. 7), additional results (Sec. 8) and an accompanying video featuring a plethora of control results (including explicit control examples of human faces, paintings and dogs). **We recommend to watch the supplementary video on our project page.**

## 2. Dog generation

In this experiment we aim to control the identity and pose of dogs, to demonstrate our capability to generalize to additional domains other than human faces.

**Implementation details:** We use AFHQ [6], 5,239 dog face images downsampled to 512x512 resolution. We use StyleGAN2 with non-leaking augmentation [14] as before, for paintings. For preserving dog IDs we use two different models ($M^0_{ID}$, $M^1_{ID}$). $M^0_{ID}$ is a dog-face recognition model DogFaceNet [18]. We noticed that DogFaceNet often recognizes two dogs to be the same if a similar pattern is present on their heads, even if they are of different breeds. To overcome this issue, we use ResNet18 [12] trained on ImageNet [7] as $M^1_{ID}$ (the penultimate layer is used for the distance comparison). We use this model since of the 1K classes in the ImageNet challenge, 90 are dog-breeds. For $M_{pose}$ we use the same pose detection network as for human faces and paintings [20].

**Photorealism:** The FID scores are 8.74 and 8.22 for our controlled and for the baseline models, respectively.

**Qualitative evaluation:** Fig. 1 shows precise control over dog pose using $E_{pose}$. We show that the pose estimation network was able to provide enough guidance to the GAN even though the domain gap between dog and human faces is quite large. Since dogs are rarely photographed from below, our model was not able to generate images of dogs with a positive pitch. Unlike humans, dogs exhibit a large variability in head-roll, therefore the model was able to capture this aspect well.

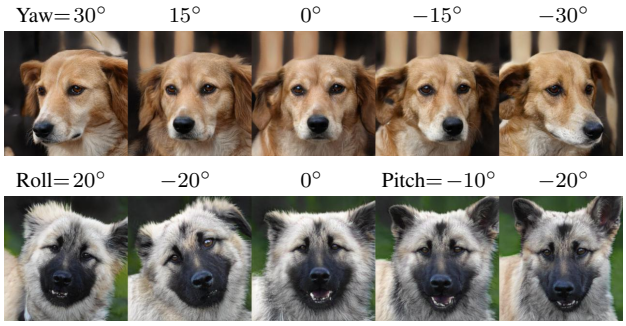


Figure 1: **Controlling head pose in dog images:** Generation results using $E_{pose}$.

## 3. Implementation details

In this section we provide further implementation details of our method and additional details on our explicit control experiment.

### 3.1. GAN-control implementation

Table 1 shows the dimensions of our latent sub-vectors $\mathbf{z}^k$, $\mathbf{w}^k$ and the dimensions of our per attribute control inputs $y^k$. Our GANs are based on the StyleGAN2 [16][1] framework. Next we list our contrastive loss components, $l_k$, and the corresponding models, $M_k$, used to compute each component:

**Face generation:**

- **ID:** ArcFace [8][2] embedding vector outputs are compared using the cosine distance.
- **Pose:** HopeNet [20][3] yaw, pitch and roll outputs are compared using $L_1$.

[1] https://github.com/rosinality/stylegan2-pytorch
[2] https://github.com/TreB1eN/InsightFace_Pytorch
[3] https://github.com/natanielruiz/deep-head-pose

| Attribute | dim $\mathbf{z}^k$ | dim $\mathbf{w}^k$ | dim $y^k$ | Description of $y^k$ |
|---|---|---|---|---|
| Faces | | | | |
| ID | 128 | 128 | | |
| Pose | 64 | 64 | 3 | yaw, pitch and roll |
| Exp. | 64 | 64 | 64 | $\beta$ of 3DMM |
| Illum. | 64 | 64 | 27 | $\gamma$ of 3DMM |
| Age | 64 | 64 | 1 | yo |
| Hair c. | 64 | 64 | 3 | Average RGB value |
| Other | 64 | 64 | | |
| Paintings | | | | |
| ID | 128 | 128 | | |
| Pose | 64 | 64 | 3 | yaw, pitch and roll |
| Exp. | 64 | 64 | 64 | $\beta$ of 3DMM |
| Age | 64 | 64 | 1 | yo |
| Style | 128 | 128 | | |
| Other | 64 | 64 | | |
| Dogs | | | | |
| ID | 192 | 192 | | |
| Pose | 192 | 192 | 3 | yaw, pitch and roll |
| Other | 128 | 128 | | |

Table 1: **Dimensionalities of latent sub-vectors and control inputs:** the table shows the dimensions of $\mathbf{z}^k$, $\mathbf{w}^k$ and $y^k$ for each of the attributes $k$.

- **Expression:** each model output of the ESR [22][4] ensemble is concatenated to a vector and compared using $L_1$.
- **Illumination:** R-Net [10][5] $\gamma$ output is compared using $L_1$.
- **Age:** Dex [19][6] outputs are compared using $L_1$.
- **Hair color:** We compute masks for the hair regions of the generated images using PSPNet [23][7]. Then, the average RGB color of each hair region is calculated to be compared using $L_1$. For hair regions with less pixels then a certain threshold the loss is not calculated, taking into account the possibility that the person in the image is bald.

**Painting generation:**

- **ID, pose, expression, age:** same as for face generation.
- **Artistic style:** we use Gatys' *et al.* [11] and Johnson's *et al.* [13] style loss. Intermediate layers of VGG16 [21] are extracted, their Gram matrices are calculated, flattened and compared via $L_2$.

[4] https://github.com/siqueira-hc/Efficient-Facial-Feature-Learning-with-Wide-Ensemble-based-Convolutional-Neural-Networks

[5] https://github.com/microsoft/Deep3DFaceReconstruction

[6] https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

[7] https://github.com/YBIGTA/pytorch-hair-segmentation

**Dog generation:**

- **ID$_0$:** DogFaceNet [18][8] outputs are compared using $L_1$.
- **ID$_1$:** the penultimate layer of ResNet18 [12], trained for ImageNet [7] classification is compared using $L_1$.
- **Pose:** same as for face generation.

For the second training phase of all the domains, the attribute values $y_i^k$ for the datasets $\{\{\mathbf{w}_i^k, y_i^k\}_{i=1}^{N_s}\}_{k=1}^{N}$ are computed using the above models, $M_k$, with the exception that the expression attribute values $y_i^{exp}$ are computed using the $\beta$ output of R-Net [10] corresponding to the expression coefficients of the 3DMM [5].

### 3.2. Explicit control analysis implementation

In Section 4.1 of the main submission, we have performed quantitative analysis of the per-attribute control precision, comparing our approach with DFG [9] and CONFIG [17]. In this Section we add some implementational details on how we made the quantitative analysis comparable between the methods.

- The age control precision is reported only for our method as it is not handled by the other two methods.
- CONFIG's expression, illumination and hair color controls are semantically different from Ours, and therefore are not suitable for comparison. Moreover, as the expression and illumination have different dimensionalities, this further complicates direct comparison.
- Ours' and CONFIG's pose were predicted using HopeNet [20]. DFG's pose was predicted using the R-Net's [10] output $\theta$, converted to degrees. This is done for the sake of comparison fairness, in order not to degrade DFG's performance as a result of using a non-compatible pose estimation model (DFG generates images aligned for R-Net).
- All methods used the truncation trick with $\psi = 0.7$ (Ours and DFG used the attribute-preserving truncation trick proposed in [9]).

[8] https://github.com/GuillaumeMougeot/DogFaceNet


Figure 2: **Qualitative control-by-example ablation study**

# 4. Analysis

## 4.1. Qualitative control-by-example ablation study

In Fig. 2 we sample three *source* latent vectors and four *target* latent vectors. We modify the source images by substituting incrementally growing subsets of their sub-vectors, $\mathbf{w}^k$, with the corresponding ones from the target images. We then validate that, perceptually, the resulting image corresponds to the correct combination of the source (unmodified) and the target (modified) attributes. From the Fig. 2 we observe that all of our unmodified controlled attributes are well preserved and the modified attributes correspond to the target. For example, looking at the Source #2 and Target #3 at Fig. 2(a), note that the person's expression (smiling) is preserved while the head orientation changed accordingly (looking to the right). As expected, looking at the same source-target pair at Fig. 2(b), note that both the source's head orientation and expression changed accordingly (looking to the right and sad). We also note that some correlations of non-controlled attributes within the dataset are partially encoded in the GAN's latent space. For example, in Source #2, the background changes as we modify the illumination (Fig. 2(c) vs (d)). This is expected as images with high-intensity illumination usually occur outdoors, while low-intensity illumination may occur indoors or at night. Anecdotally, in Source #2 and Target #2 a microphone appears when introducing a more serious expression.

## 4.2. Disentangled projection ablation study

In this section we provide a qualitative ablation study of the proposed method for disentangled projection. First, we project the images to the disentangled latent space of

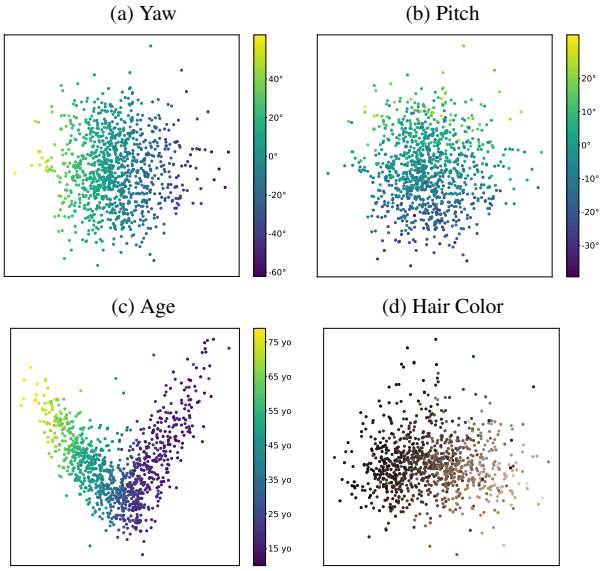


Figure 3: **PCA of embedding space:** we present a visualization of the sub-vector spaces ($\mathbf{w}^{pose}$, $\mathbf{w}^{age}$ and $\mathbf{w}^{hair}$) for 10K samples projected to their corresponding truncated two-dimensional PCA sub-spaces. The color-coding of points in (a) and (b) correspond to the predicted values of yaw and pitch in degrees, respectively. It appears that the first principal component corresponds to yaw while the second corresponds to pitch. The color-coding in (c) corresponds to the predicted age in years. The first principal component of the age latent space is enough to explain 64% of the variation. The color-coding in (d) represents the estimated average hair color for each point in the latent space. It appears that there is a rough correlation between hue and saturation to the first and second principal components, respectively.

a trained GAN. Second, we modify the sub-vectors associated with pose, illumination and age. We show four types of projection methods: (a) naïve projection to latent space $\mathcal{W}$, (b) projection to the extended latent space $\mathcal{W}^+$ [4], (c) projection to a partially extended latent space, where we only extend the sub-spaces associated with ID and other $\mathcal{W}^+_{ID,other}$, and (d) the same projection as (c) with an additional constraint that the remaining sub-vectors reside on approximated linear sub-spaces of their corresponding manifolds. We achieve the above using the following approach: we perform PCA for each latent subspace of 10K randomly sampled sub-vectors $\mathbf{w}$, where the number of components are selected so as to preserve 50% of the variance. In practice, this number is very low. For expression, pose, illumination and hair - two components. For age only one component. Visualizations of these spaces are presented in Fig. 3. During the optimization process, after each gradient descent step, we linearly project the latent sub-vectors to the truncated PCA spaces and re-project them back to their corresponding spaces. We present visual results of selected images in Fig. 4. The first column shows the real input image. The second column shows the images that were created by the GAN using the projected latent vectors. The rest of the columns show the images generated using the projected latent vector and explicitly controlling the pose, illumination or age sub-vectors via $E_{pose}$, $E_{illum}$ or $E_{age}$. Images in the first row (projection method a) suffer from inaccurate reconstruction, identity loss and severe artifacts (for an extreme example, see third person). The images in the second row (projection method b) exhibit a significantly better reconstruction. However, when modifying the sub-vectors of other attributes, the image's quality deteriorates and strong artifacts are visible. This is prominently exhibited in the unnatural colors in the images of the first person. The last two rows of each person (projection methods c and d) demonstrate that using the techniques described above these artifacts are substantially mitigated, leading to images with high reconstruction accuracy, that preserve the identity and remain artifact-free.

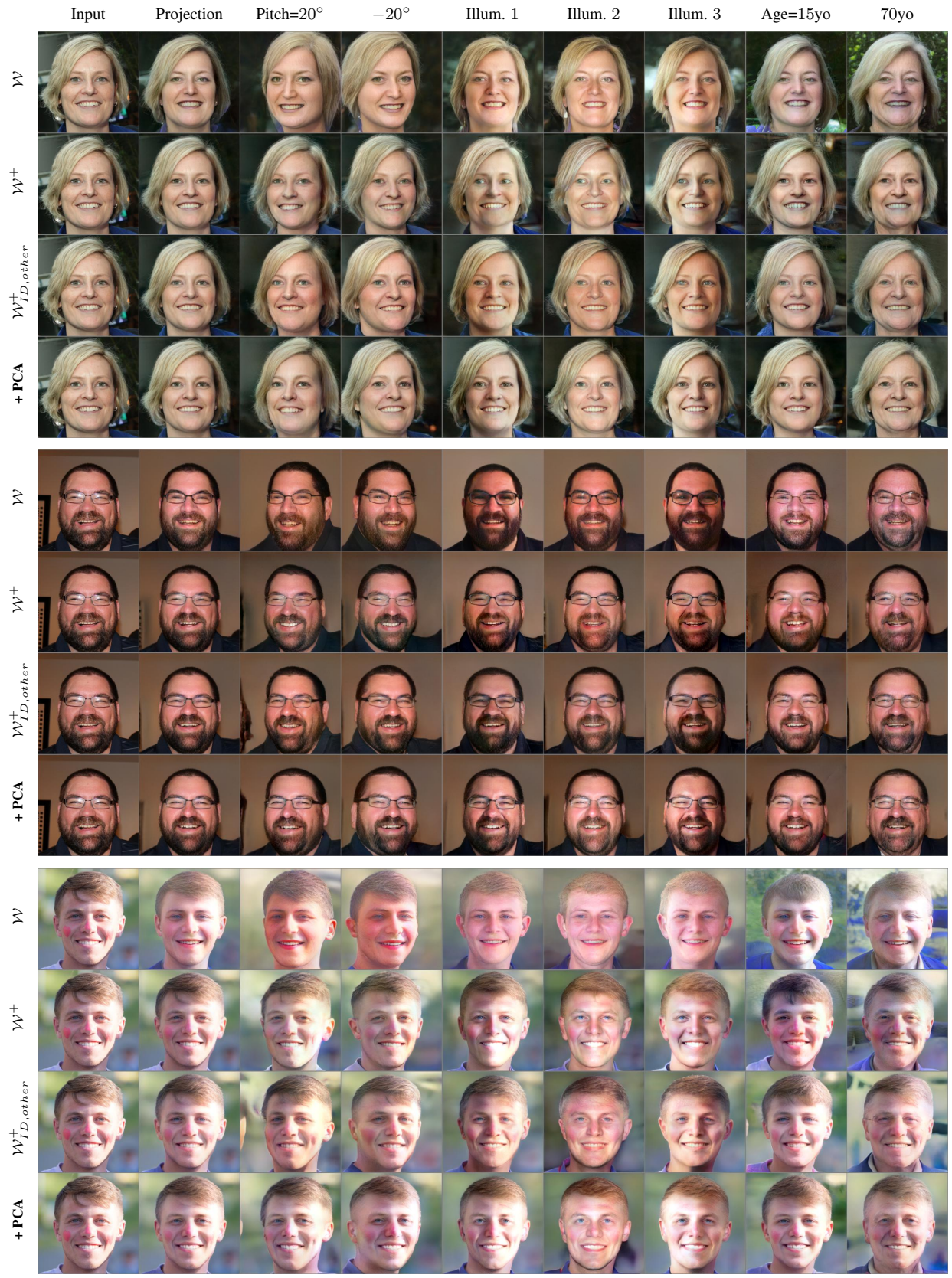


Figure 4: **Disentangled projection ablation study [1, 2, 3].**

CONFIG [17] (with truncation trick $\psi = 0.7$)

DFG [9] (with truncation trick $\psi = 0.7$)

**Ours** (with truncation trick $\psi = 0.7$)

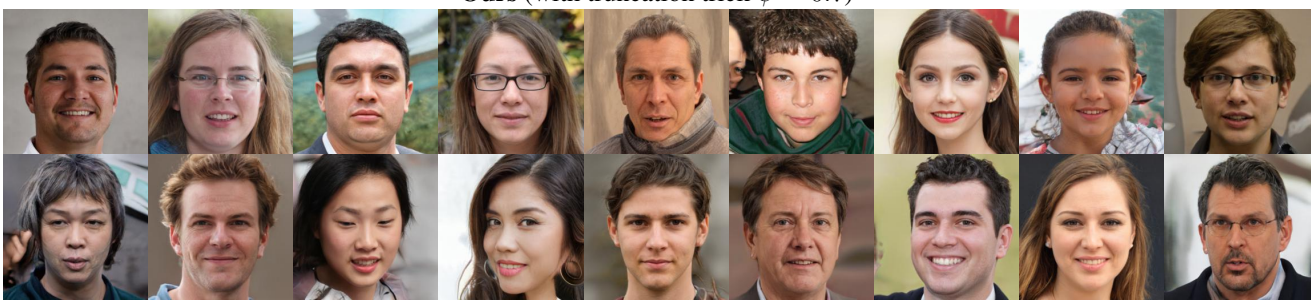

Figure 5: **Qualitative comparison:** We generated a random batch of 18 images, for each one of the methods. Rows 1-2, CONFIG [17] with the default settings for the truncation trick ($\psi = 0.7$). Rows 3-4, DFG [9] with the default settings for the truncation trick ($\psi = 0.7$). Rows 5-6, Our results with the truncation trick set to $\psi = 0.7$.

## 5. Qualitative comparison

In addition to the quantitative photorealism comparisons conducted, we further demonstrate the image quality differences between all methods with qualitative examples. Fig. 5 presents images generated by CONFIG, DFG and our method. We use the truncation trick with $\psi = 0.7$ for all generations. Ours and DFG use the attribute-preserving truncation trick.

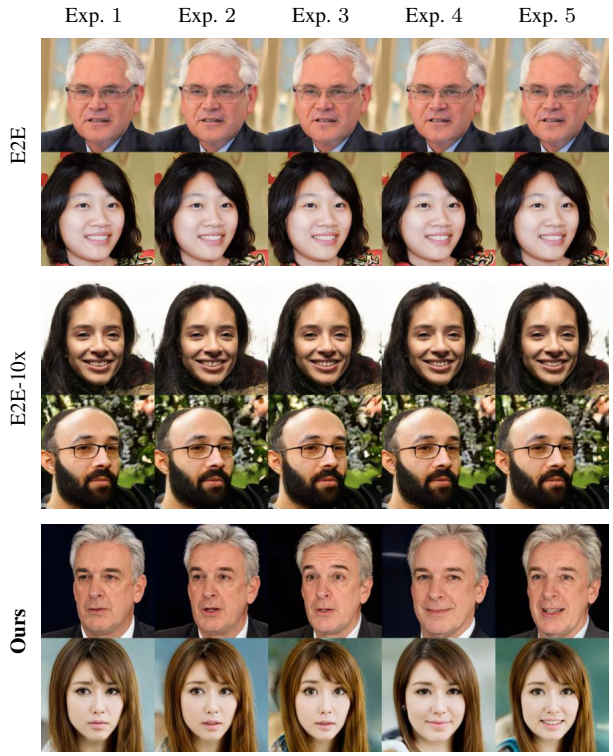


Figure 6: **Ours vs. end-to-end expression control**

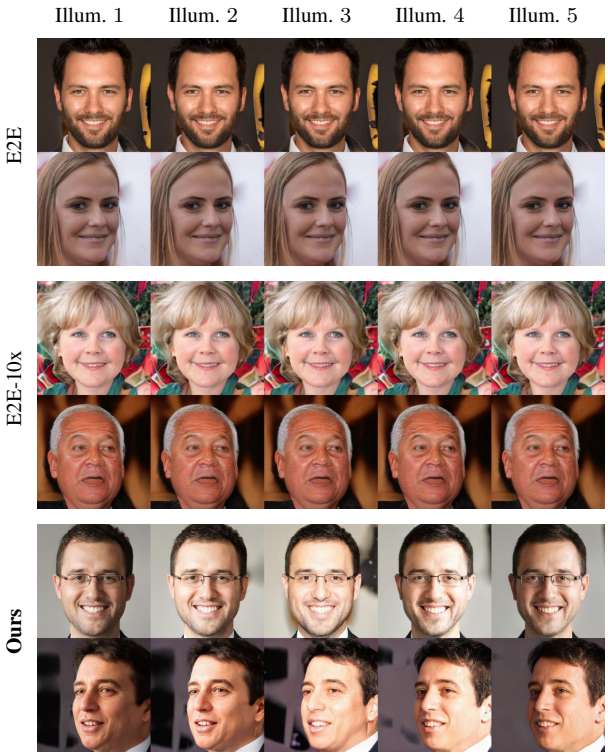


Figure 7: **Ours vs. end-to-end illumination control**

## 6. Ablation study - additional details

In this section we provide implementation details and additional qualitative comparisons for the ablation study.

### 6.1. E2E qualitative comparison and implementation details

The end-to-end model is trained in a single training phase. The architecture of the end-to-end model is the same as the final architecture we use for inference after the second training phase of our original approach (see architecture under *inference* in Figure 2 of the main paper). We derive a feasible set of input attribute control values based on the ones inferred from the actual samples in the FFHQ dataset. Practically for each image in the FFHQ dataset (70K images) we extract all its attributes via the pre-trained predictors, resulting in five attribute datasets of size 70K each (corresponding to pose, expression, illumination, age and hair color). We experimented with several options of training such an end-to-end approach. The options vary in the different configurations of the attribute matching loss coefficients and scheduling mechanisms (all training runs that used the attribute matching loss starting from the first iteration diverged after a few iterations). In the paper we present results of two models trained end-to-end, both used the matching loss starting from the 20K's iteration (when the generated images start to resemble faces). Figures 6, 7, 8 and 9, 10 show qualitative comparisons of our approach vs. E2E. This comparison is in agreement with the quantitative results presented in the main paper where the control precision and image quality are inferior to our approach.

### 6.2. NoDis implementation details

The motivation for our design choices are simplicity and fair comparison. The encoder ($E$) architecture consists of: $N$ per-attribute encoders and an ID encoder, concatenation and an aggregating encoder. First, each attribute is embedded using a per-attribute encoder. Then, all sub-encoders' output vectors are concatenated and passed through a single aggregating encoder that outputs the StyleGAN2's $\mathcal{W}$ latent vector. When training, the inputs to $E$ are the predicted attributes and the ID (predicted ArcFace embedding vector).


Figure 8: **Ours vs. end-to-end hair color control**


Figure 9: **Ours vs. end-to-end pose control**

## 7. Limitations

Next, we elaborate on some of the limitations that we encountered during our experimentation. The leftmost columns of Fig. 11 shows examples in which the originally generated ID is of a child. In some of these cases, increasing the age does not necessarily translate into naturally maturing process of the person's head, but into appearance of wrinkles and other skin deformation artifacts. We hypothesize that this happens due to the limitations of our age prediction network, which was not intended to address ages below 15 and to the fact that face geometry mostly does not change after the age of 15, but does change significantly for younger ages.

Fig. 12 shows limitations of our pose control. For some controlled pose angles, severe artifacts appear in the generated images. This is to be expected as these poses are out of the distribution of poses appearing in the FFHQ dataset.

In Fig. 13 we show that in some cases of older men, the hair color is controlled by the age rather then by input hair color. We attribute this to the strong biases present in the FFHQ data set, related to the whitening hair of elderly men.

Another limitation we have noticed is that our expression control does support closed eyes or asymmetric eyebrows. This might be due to the insufficient representation of such faces in the FFHQ datasets.

In Fig. 14 we show that the IDs of the generated dogs are not preserved well when pitch is modified. We provide two possible explanations: (a) the correlation between dog size and pitch in the dataset, *i.e.*, small breeds are usually photographed at even or positive pitch, while large breeds tend to be photographed from above; and (b) the limited capacity of DogFaceNet, the recognition model used when training the GAN. The model was trained on a relatively small dataset of 8,363 images of 1,393 dogs.

Our approach does not rely on an ad-hoc facial prior such as the 3DMM. Instead, it relies on deep models trained on similar but different domains than that used to train our GAN (see Sec. 3 for details). In some cases we observe that such an approach successfully overcomes the domain gap, as with the case of the head-pose estimation model trained on the domain of human faces was accurate when applied to dogs and paintings, while in other cases, such as DogFaceNet, the model's accuracy was insufficient.

To conclude, we believe that much of our model's limitations are either caused by inherent biases in the datasets or are due to limitations of the pre-trained models on which our method relies.


Figure 10: **Ours vs. end-to-end age control**

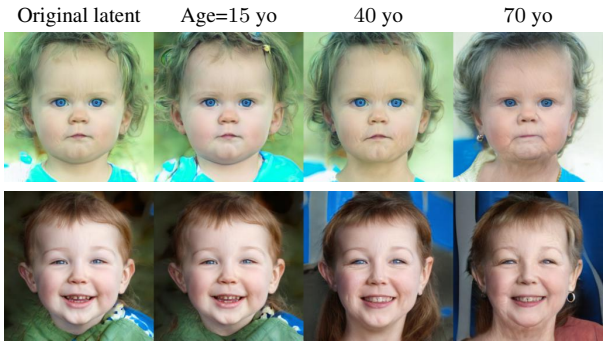


Figure 11: **Age limitations:** We observed that for some unmodified latent vectors, $\mathbf{w}$, which were used to produce images of young children or infants, the age control tends to provide unrealistically or not compatible with the input age, results. The leftmost column shows results of the unmodified latent vectors and the other columns are controlled results using $E_{age}$ with the above inputs.

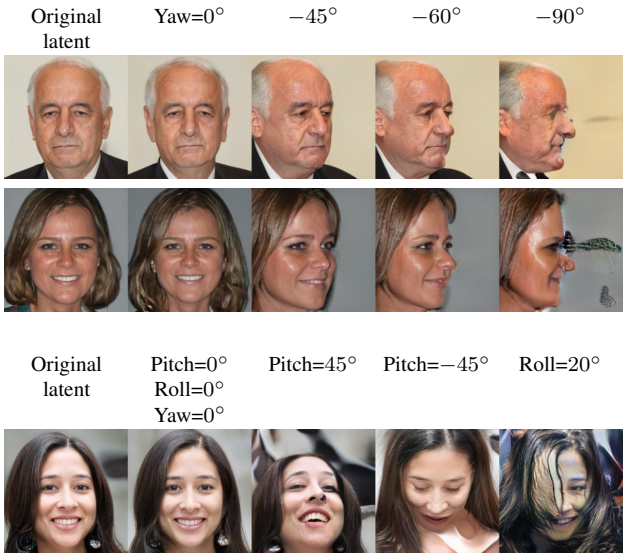


Figure 12: **Pose limitations:** We observe appearance of severe artifacts when the pose angles are out of the range of possible angles appearing in the FFHQ dataset [15]. Rows 1-2 show deterioration of the generated images when yaw is set lower then $-45°$. Row 3 shows that for extreme pitch angels the image deteriorates and that for roll angels the image gets corrupted. This is reasonable as there are no images in FFHQ with a noticeable roll.

## 8. Additional Results

Figures 15, 16, 17, 18 and 19 show additional results for explicitly controlling human face pose, illumination, expression, age and hair color, respectively. Figures 20 and 21 show additional results of controlling paintings age and pose, respectively. Fig. 22 shows additional results for controlling dogs' poses.

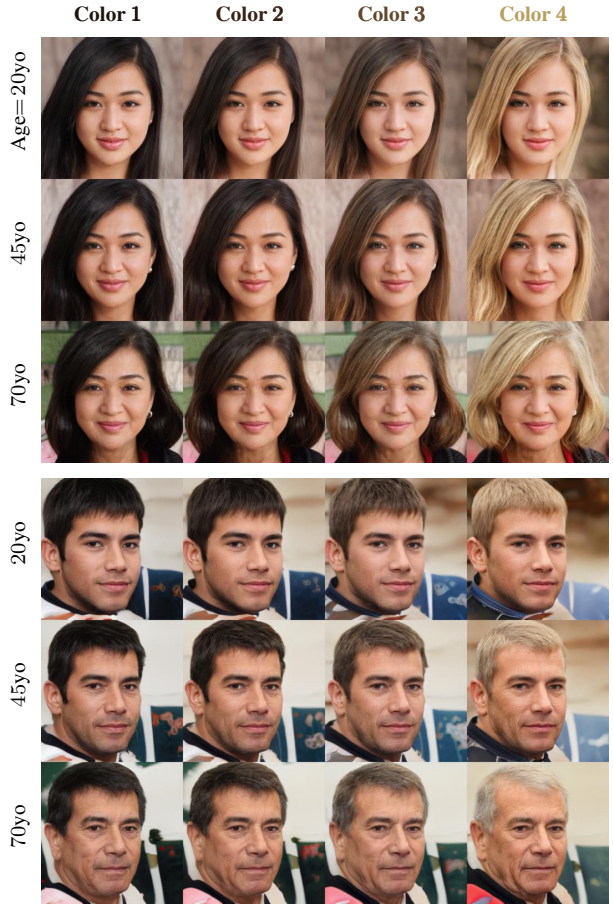


Figure 13: **Hair color limitations:** We observed that for some generated images the hair color is not solely depended on $\mathbf{w}^{hair}$ as expected. This happens mostly for generated images of older men. As can be seen in rows 4-6, for some hair color inputs (mostly bright ones), $y^{hair}$, when the age increases the hair whitens.

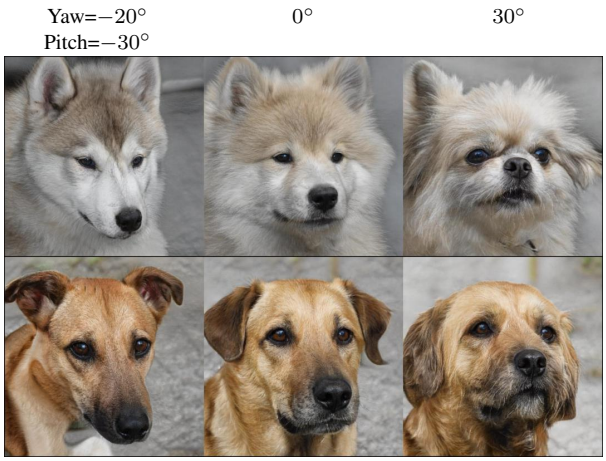


Figure 14: **Dog pitch limitations:** We noticed that changing the dog's pitch is effecting the dog's identity (or breed). This may be due to the small size of the dataset and due to the fact that small dogs tend to be photographed at face level rather than from above.


Figure 15: **Controlling pose**


Figure 16: **Controlling illumination**


Figure 17: **Controlling expression**


Figure 18: **Controlling age**


Figure 19: **Controlling hair color**


Figure 20: **Controlling paintings age**

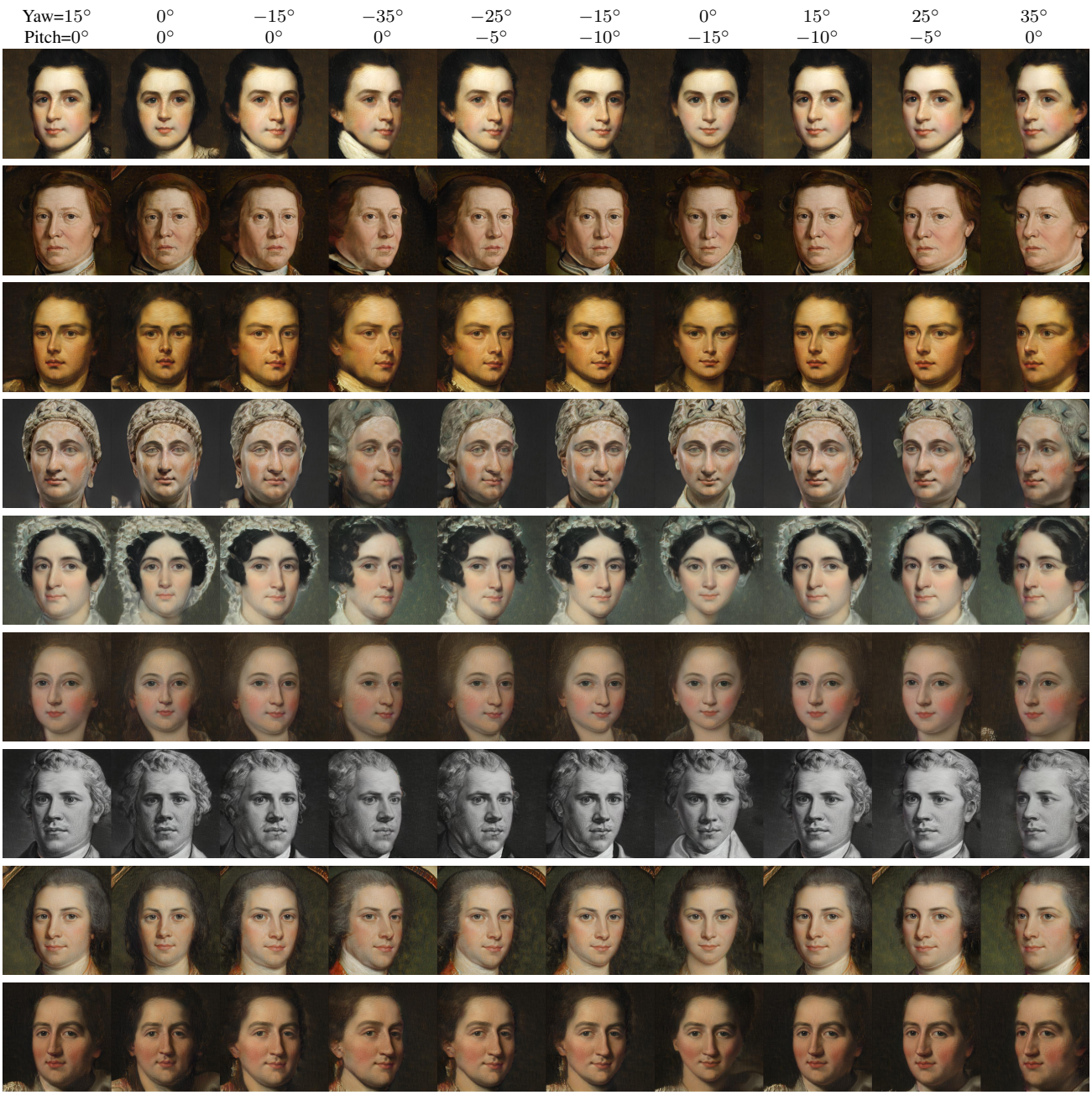


Figure 21: **Controlling paintings pose**

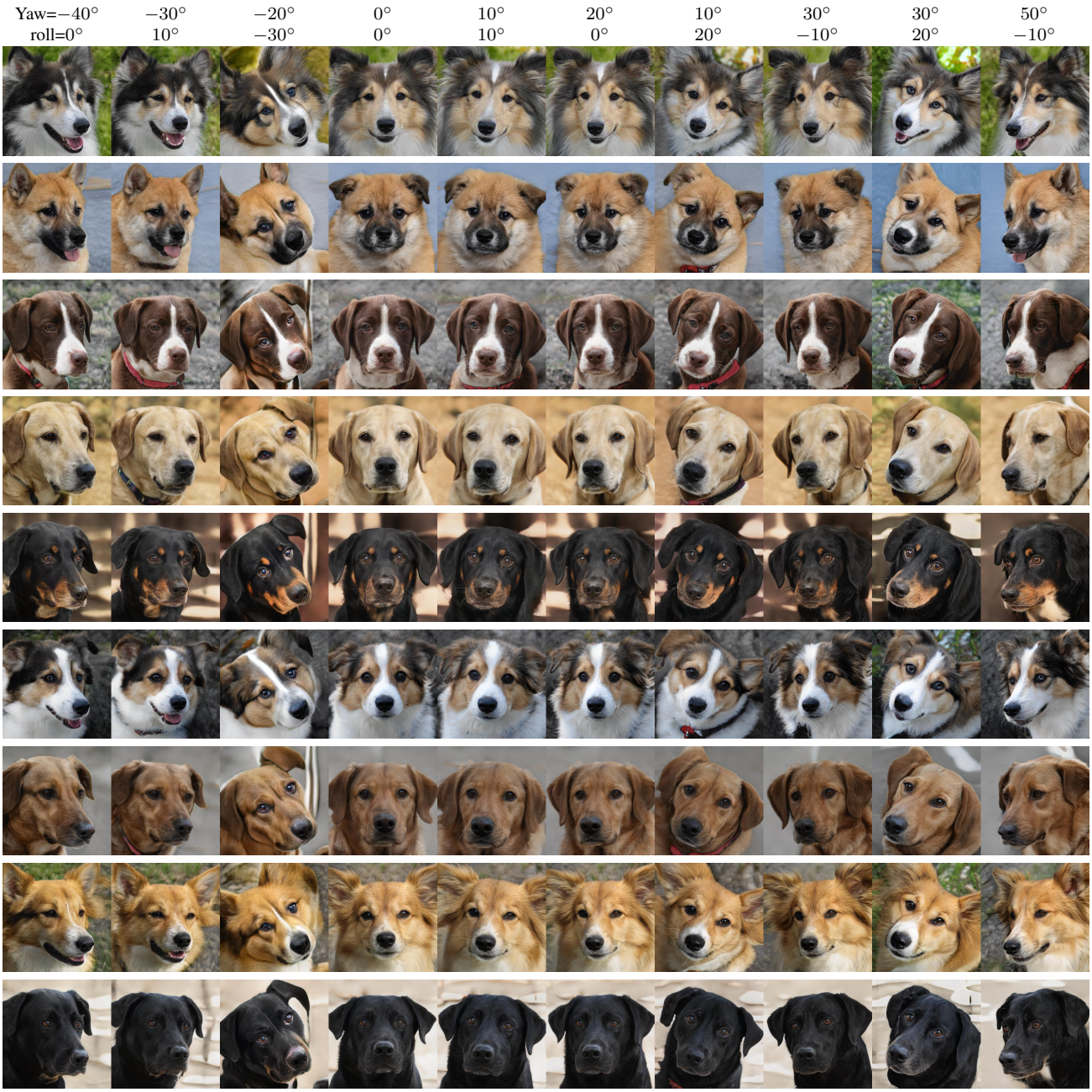


Figure 22: **Controlling head pose in dog images**

## References


[1] The original image is at http://www.flickr.com/photos/cfccreates/8577115747 and is licensed under: http://www.creativecommons.org/licenses/by/2.0. 5

[2] The original image is at http://www.flickr.com/photos/onionboy/7332990776 and is licensed under: http://www.creativecommons.org/licenses/by-nc/2.0. 5

[3] The original image is at http://www.flickr.com/photos/62487011@N08/28696918755 and is licensed under: http://www.creativecommons.org/licenses/by-nc/2.0. 5

[4] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to Embed Images Into the StyleGAN Latent Space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 4

[5] Volker Blanz and Thomas Vetter. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999. 2

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2020. 1

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1, 2

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1

[9] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning. In *IEEE Computer Vision and Pattern Recognition*, 2020. 2, 6

[10] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D Face Reconstruction with Weakly-Supervised Learning: from Single Image to Image Set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 2

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image Style Transfer Using Convolutional Neural Networks. In *CVPR*, 2016. 2

[12] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 2

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[14] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 1

[15] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 4401–4410, 2019. 9

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. *CoRR*, abs/1912.04958, 2019. 1

[17] Marek Kowalski, Stephan J. Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. CONFIG: Controllable Neural Face Image Generation. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 6

[18] Guillaume Mougeot, Dewei Li, and Shuai Jia. A Deep Learning Approach for Dog Face Verification and Recognition. In *PRICAI 2019: Trends in Artificial Intelligence*, pages 418–430, Cham, 2019. Springer International Publishing. 1, 2

[19] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep Expectation of Apparent Age from a Single Image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015. 2

[20] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-Grained Head Pose Estimation Without Keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 1, 2

[21] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015. 2

[22] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient Facial Feature Learning with Wide Ensemble-based Convolutional Neural Networks, Feb 2020. 2

[23] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *CVPR*, 2017. 2