Expressive Talking Head Video Encoding in StyleGAN2 Latent Space – Supplementary –

Trevine Oorloff Yaser Yacoob
University of Maryland, USA
{trevine,yaser}@umd.edu

A. Overview

The outline of this document is as follows.

- Sec. B: Detailed steps on alignment in the pre-processing stage
- Sec. C: Discussion on the GAN inversion stage
- Sec. D: Additional details on identity-latent selection
- Sec. E: Illustrated explanations of noteworthy sections of facial attribute encoding
- Sec. F: Further details and examples of experiments, results, and limitations of the proposed framework
- Sec. G: Discussion on Potential Negative Societal Impact

Additionally, please refer to the supplementary video, for qualitative examples and a brief overview of the proposed approach.

B. Video Pre-Processing: Alignment

The alignment carried out in the pre-processing stage could be elaborated further using the three steps below.

1. Detect eye blinking and compensate for its effect on landmark location of the eyes. This improves StyleGAN2-based alignment by removing the sensitivity to eye shape change during blinking.
2. Registration of the face between a frame and a key frame uses a parameterized affine optical-flow model of the head [2], excluding the non-rigid face features (eyebrows, eyes, and mouth). The over-constrained optical-flow model is very effective at stabilizing the face between consecutive frames unless there are changes in the Yaw/Pitch of the head. We employ the mean L2 distance to automatically determine the quality of the inter-frame alignment over the non-rigid parts of the face (*i.e.*, compute the residual error in RGB values of face stabilization). A mean distance beyond a fixed threshold indicates that the affine motion model is not successful at stabilizing the rigid part of the face, triggering step (3).
3. Key frame change that forces a new key frame to be the basis for future frames' face stabilization (aligned according to step (1)).

For optical flow head registration, the threshold of the mean RGB registration error over the face (excluding eyes, mouth, and eyebrow areas) had to exceed 45.0 (if the inter-frame Yaw and Pitch change is less than 2°), or 30.0 (if the inter-frame Yaw or Pitch change exceeds 2°). The objective is to avoid forcing face registration when the head is moving out-of-plane. Instead, a change in the key frame is triggered, allowing the StyleGAN2 encoder to capture the new head-pose. Fig. 1 depicts the key frames from a short sequence when the head moves to near profile and then back.

C. GAN Inversion

Two factors were considered in choosing an appropriate GAN inversion method: (1) faithful representation of the given image (*i.e.*, minimal reconstruction loss), (2) ability to facilitate latent space edits. The authors of [14] suggested that there exists a trade-off between these two factors, *i.e.*, distortion and editability. Generally, inversion is done using a trained encoder and/or an optimization framework. While the former has better editability, it has a comparatively high reconstruction loss and vice versa. We chose the e4e encoder [14], which was designed to facilitate the inversion of real images in proximity to the regions StyleGAN2 was trained on, thus mitigating the trade-off.

The e4e encoder while producing state-of-the-art results in GAN inversion of real images, has a few failing instances. For certain subjects, (*e.g.*, Fig. 2 (a)) the identity of the encoded image deviates considerably from the real frame. In such cases, as we perform the inversion per-frame, there is a tendency for the identity to change across the frames of a single video clip as well. The identity change across

Figure 1. **Key frames in a video sequence with head out-of-plane rotation**

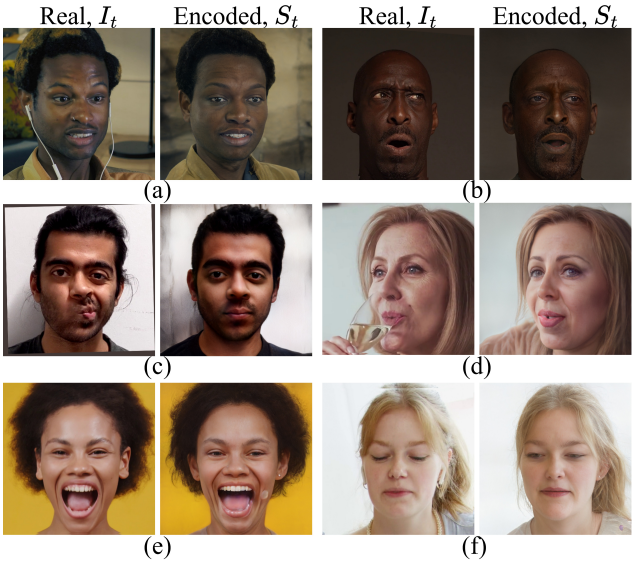

Figure 2. **Few examples of isolated instances where the e4e encoder fails**. The examples depict: (a) poor identity, (b) incorrect gaze, (c) inability to capture extreme mouth movements, (d) deformations caused due to occlusions, (e) visual artifacts, and (f) flaws in facial features captured (open eyes while closed in real)

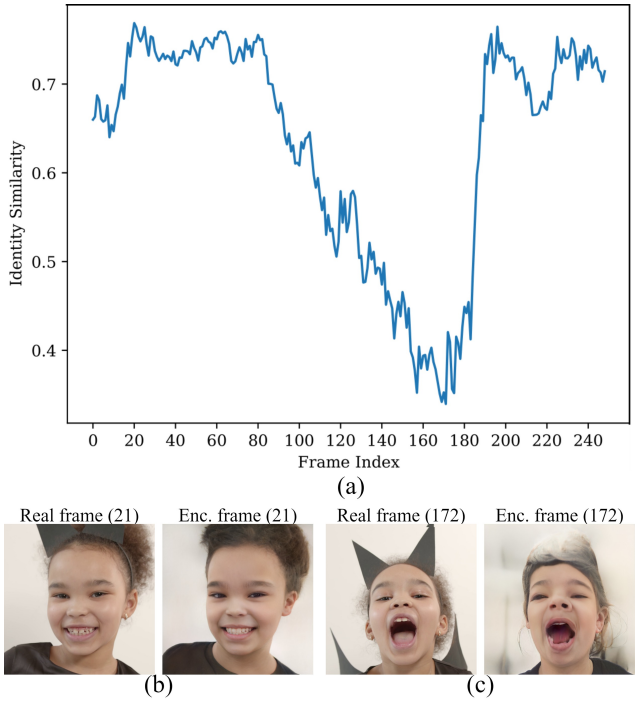

Figure 3. **Identifying L_{ID} is based on identity matching using ArcFace** [3]. (a) depicts the identity similarity scores computed between the encoded and real frames. In this case, (b) the best identity is at frame 21 while (c) the worst is at frame 172

frames could be due to the poor convergence of the encoder resulting from the existence of a higher per-frame loss due to poor identity. Additionally, there exist cases where the e4e encoder fails to capture certain facial attributes successfully (*e.g.*, Fig. 2 (b) and (c)) which could be due to the low representation of complex features in the StyleGAN2 training dataset (FFHQ [8]). Further, certain visual artifacts and deformations tend to appear in certain cases similar to the examples shown in Fig. 2 (d), (e), and (f), which could be caused due to occlusions (d) and the noisiness in the neighborhood of the inverted $W+$.

However, the impact of most of these issues on the resynthesis is mitigated as (1) we anchor our deformations with respect to a single ID-frame that has the highest identity match with the real and (2) utilize PTI [13] to minimize the disparity between the real and synthesized frames.

D. Identity-Latent Selection

The per-frame inversion creates a series of latents. Depending on the extent of head motion, deformation in StyleGAN2 space is likely. Therefore, the choice of the ID-frame is of great significance as it serves as the base identity for the face and head-pose deformations across the entire sequence of frames. Hence, we first use ArcFace [3] to compute the similarity between the source and the reconstructed images of the face and then the (1) closest of the face-matches that

Facial Attribute, $\mathcal{F}$	StyleSpace Indices, $\mathcal{V}$
Mouth	{**6:** 113, 202, 214, 259, 378, 501}, {**11:** 6, 41, 78, 86, 313, 361, 365}, {**8:** 17, 387}, {**14:** 12},{**15:** 45}
Chin/ Jaw	{**5:** 50, 505}, {**6:** 131}, {**8:** 390}
Eyes	{**9:** 63}, {**11:** 257}, {**12:** 82, 414}, {**14:** 239}, {**17:** 28}
Eyebrows	{**8:** 6, 28}, {**9:** 30}, {**11:** 320}
Gaze	{**9:** 409}

Table 1. **StyleSpace indices corresponding to the deformation of facial attributes**. The indices take the form of $\{l\colon c_1, c_2, \dots\}$, where l and c denote the layer index and channel index of the StyleSpace

is also (2) near frontal view of the person, and (3) has no blink is chosen as the representative L_{ID}, the basis for re-synthesis. An example plot depicting the variation of the identity similarity (computed based on ArcFace) is given in Fig. 3 (a) and the corresponding best and worst ID-frame candidates based on our criteria are shown in Fig. 3 (b).

E. Facial Attribute Encoding

E.1. Head-Pose Encoding

The flow of the head-pose encoding is illustrated in Fig. 4. Moreover, to evaluate the significance of our optimization based head-pose encoding approach, we compare our results post head-pose adjustment against the straightforward use of StyleFlow with the $\{Y_t, P_t\}$ parameters computed using [1]. While quantitative results on 5 sample videos were provided in Tab. 3 of the main paper, please refer to the supplementary video for qualitative comparisons. It could be seen that our approach captures the head-pose well and has a significantly low jitter compared to the straightforward approach with StyleFlow.

E.2. Choice of StyleSpace Indices

We illustrate the facial deformations corresponding to the manipulation of each of the 32 StyleSpace indices tabulated in Tab. 1 in Fig. 8. A pair of images marked as $(l, c) : +/-$ is included for each StyleSpace index, $(l, c) \in \mathcal{V}$ denoting the sign of the perturbation added to the respective StyleSpace index.

E.3. Index Specific Learning Rate

The variation of index sensitivity computed over the indices corresponding to the {mouth + chin/jaw} is shown in Fig. 5 (a). The significant variations seen in the plot make it evident that the index sensitivities cannot be simply ignored and hence, the indices cannot be treated the same during optimization. In order to alleviate the dominance of indices with a higher index sensitivity, we compute an index specific learning rate, $\eta_{f,i}$, $\Gamma_{f,i}$ specified in Eq. (5) in the main-paper. The $\Gamma_{f,i}$ corresponding to the indices in Fig. 5 (a) are depicted in Fig. 5 (b). It could be seen that the $\eta_{f,i}$ of indices having a higher $\Gamma_{f,i}$ is comparatively lower than the indices of lower $\Gamma_{f,i}$, thus effectively alleviating the dominance.

E.4. Details on Optimization

The AdamW [10] optimizer with AMSGrad [11] was utilized with an initial learning rate of $\eta = \{\eta_{f,i}; \forall f \in \mathcal{F}, i \in \mathcal{V}\}$, $(\beta_1, \beta_2) = (0.9, 0.999)$, and $\epsilon = 1e^{-8}$. The optimization was over 100 epochs ($N = 100$) and the learning rate was decayed every 10 epochs with a decaying factor of 0.8 using a learning rate scheduler for improved convergence. The optimization takes approximately 1 min./frame on a single GTX1080Ti GPU. Additional details on the loss terms defined in equations (9) - (12) of the main-paper are given below.

$\boldsymbol{\mathcal{L}_{LPIPS}}$: The LPIPS loss [20], which is known to learn perceptual similarities well [5, 12], was used to capture the structural details of the facial attributes between S_t and $\hat{S}_t$. Nevertheless, $\mathcal{L}_{LPIPS}$ was not used in solving for the gaze ($\mathcal{L}_p$) as it is invariant to subtle spatial changes and hence introduces a slight jitter when used.

$\boldsymbol{\mathcal{L}_{L2}}$: This denotes the L2 norm between the S_t and $\hat{S}_t$, and enables precise reconstruction (*e.g.*, the case of gaze).

$\boldsymbol{\mathcal{L}_{ID}}$: To mitigate the risk of changing the identity of the subject across frames while optimizing over the latent space, the identity loss [12] is in place as a regularization term. This is computed between $\hat{S}_1$ and $\hat{S}_t$.

$\boldsymbol{\mathcal{L}_{FP}}$: As we optimize over 32 indices in parallel, we noted occasional nose, mouth, and chin/jaw deformations. To discourage unwarranted deformations, the Face-Parsing loss, which is the L2 norm of the difference between the masked face-parsing scores [18] of the rendered and target frames, is used instead of facial-landmark coordinates loss (*e.g.*, [1]). Face-parsing scores facilitate the gradient flow through the optimization and are more precise and stable across the frames.

$$\mathcal{L}_{FP} = ||FP(\hat{S}_t) * M - FP(S_t) * M||_2 \qquad (1)$$

where function $FP(\cdot)$ yields face-parsing scores and M denotes the binary mask of the face.

$\boldsymbol{\mathcal{L}_{IF}}$: The inter-frame loss is a derivation of the Frame Difference-Based (FDB) loss proposed in [17], to enforce temporal coherence between frames. We minimize this loss along with the other spatial losses to avoid enforcing temporal continuity posteriori. Provided the target video is temporally coherent, this loss is based on the concept that the image space and feature space differences between consecutive frames embed the temporal coherence. We use LPIPS and L2 losses to compute differences in the feature and im-

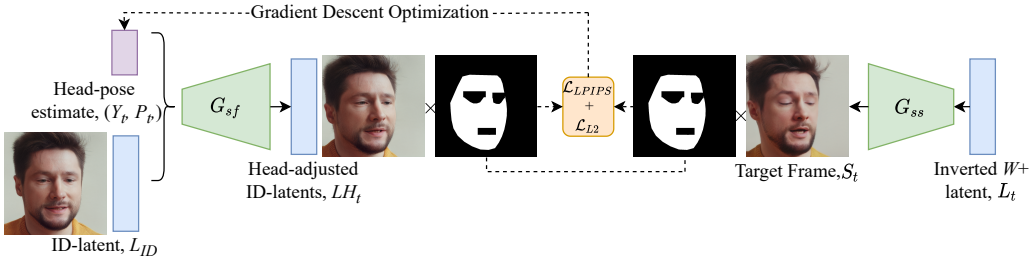

Figure 4. **The per-frame head-pose optimization flow using StyleFlow Yaw/Pitch**. We re-formulate the head-motion as a head-pose matching problem between a rendered image of the real-frame's encoded latent, L_t, and the rendered image of a rotated L_{ID} which is solved as a minimization problem employing L2 and LPIPS losses (computed over a masked area of the face excluding non-rigid areas) to search the Yaw-Pitch space using gradient descent.

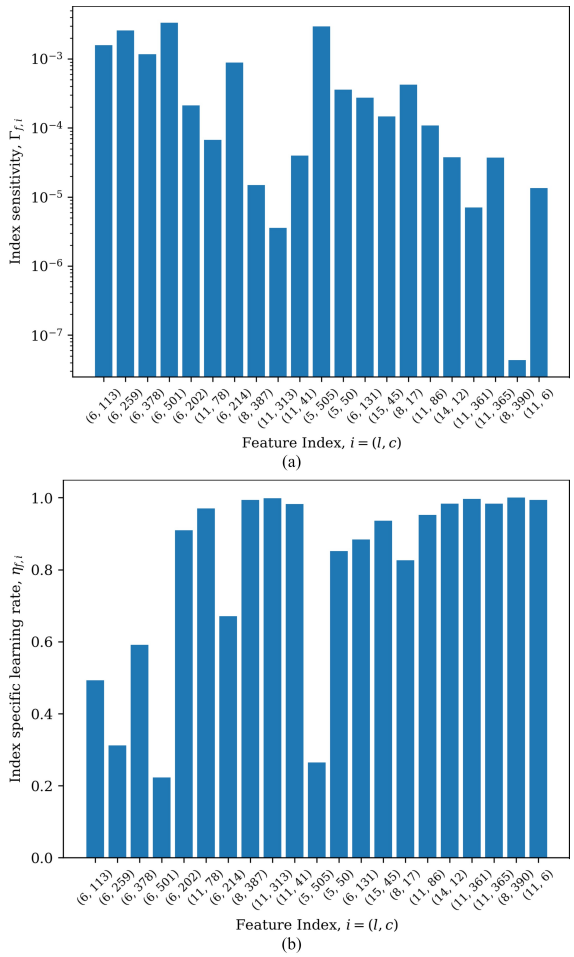

Figure 5. **(a) Index sensitivity and the corresponding (b) index specific learning rate**. This figure represents values computed for an example subject over the {mouth+chin/jaw} indices. It could be seen that the $\eta_{f,i}$ of indices having a higher $\Gamma_{f,i}$ is comparatively lower than the indices of lower $\Gamma_{f,i}$, thus effectively alleviating the dominance during the optimization.

age spaces, respectively.

$$\mathcal{L}_{IF} = \mathcal{L}_{IF\_LPIPS} + \mathcal{L}_{IF\_L2} \tag{2}$$

$$\mathcal{L}_{IF\_*} = \mathcal{L}_{*}\{S_t, S_{t-1}\} - \mathcal{L}_{*}\{\hat{S}_t, \hat{S}_{t-1}\} \tag{3}$$

where $*$ denotes either LPIPS or L2.

F. Experiments and Results

F.1. Dataset

As stated in Section 4.1 of the main-paper, we compose a dataset consisting of video clips of 4K resolution sourced from the site `www.pexels.com`. The videos were chosen such that diverse subjects belonging to various ethnicities, age groups, and having different facial geometries, performing significant head-pose movements and facial deformations (both expressions and speech) were included. The results were computed based on 150 videos chosen from the dataset, with a mean of 304 frames, a minimum of 100 frames, and a maximum of 1000 frames.

F.2. Evaluation Metrics

The following metrics were used for the quantitative evaluation of our re-enactment videos in comparison with baselines, which are tabulated in Tables 1 and 2 of the main-paper.

Mean L1-distance, L1: The per-pixel L1-distance was averaged across pixels, channels, and frames to obtain the score. The pixel values of the input images were in the range of [0,255].

Learned Perceptual Image Patch Similarity Loss, LPIPS: The metric was computed per-frame using the original implementation of [20] computed using the feature space of AlexNet [9].

Identity Loss, $\mathcal{L}_{ID}$: The identity loss was computed using,

$$\mathcal{L}_{ID} = 1 - \langle \phi(S_t), \phi(\hat{S}_t) \rangle \tag{4}$$

where ϕ represents the pretrained ArcFace network [3] and $\langle \cdot, \cdot \rangle$ denotes the cosine similarity. While in re-synthesis (Table 1 in the main-paper) the loss was computed between the synthesized frame and the real frame, for puppeteering (Table 2 in the main-paper) the loss was computed between each frame and the puppet's ID-frame.

Figure 6. **Additional examples demonstrating the versatility of our algorithm in video re-synthesis**

Peak Signal to Noise Ratio, PSNR: This was computed using the built-in function of python's scikit-image package using images having pixel values in the range [0,255].

Fréchet Inception Distance, FID: This metric, which is used to measure the photo-realism between two datasets, was computed based on the original implementation of [6] with a batch size of 100. Note: The input images are rescaled to 299×299 at the input of the inception network.

Fréchet Video Distance, FVD: The spatio-temporal perceptual score measured through FVD was computed using the original implementation of [15]. Video fragments of length 120 frames were scored with a batch size of 8 and averaged to obtain the final FVD score due to resource limitations. Note: The frames are rescaled to 224×224 by the algorithm.

Fréchet Video Distance - Mouth, FVD$_M$: Similar to FVD, with the exception of the metric being scored over the masked area of the mouth region.

Action Unit, Gaze, Pose Correlations, ρ_{AU}, ρ_{GZ}, ρ_{pose} : These metrics measure the time-series correlation be-

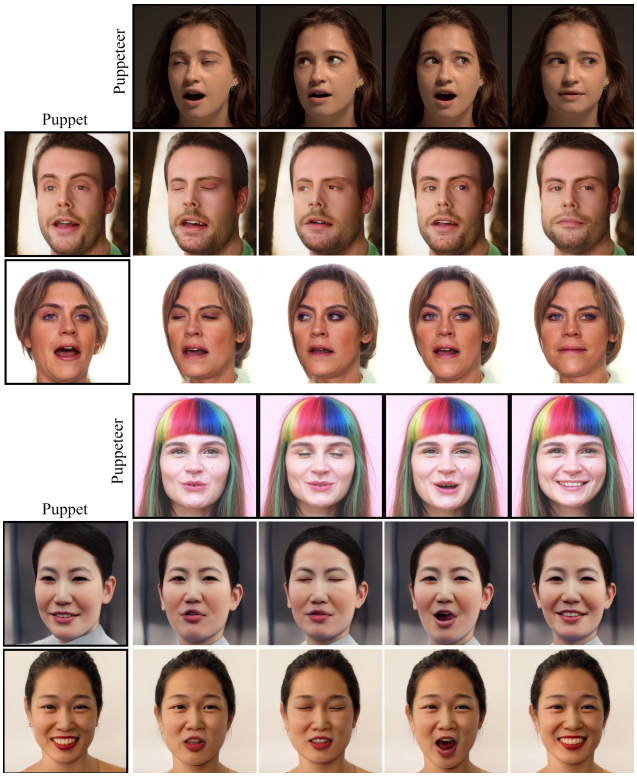

Figure 7. **Example puppeteering results** generated by applying the encoded parameters computed for the puppeteer through our encoding algorithm onto the ID-latent of the puppet

tween the Action Unit activations, Gaze angles, and Yaw and Pitch angles respectively, which are computed using OpenFace 2.0 [1] of the synthesized and the reference sequences. These provide an insight into how well the facial deformations (ρ_{AU}), eye motion (ρ_{GZ}), and pose (ρ_{pose}) are captured by the algorithm in a spatio-temporal sense.

Note: All metrics except FVD, were computed per frame and averaged across all the frames. Further, except for identity loss and correlation metrics, all other metrics were computed over a masked-out region of the reference face of each frame.

F.3. Video Results

The additional examples of video re-synthesis and puppeteering depicted in Fig. 6 and Fig. 7 respectively reaffirm the versatility of our approach. Video examples comparing the existing state-of-the-art approaches could be viewed in the supplementary video. In comparison to our results, visual artifacts, lack of sharpness, and incorrect pose and facial deformations could be observed in the re-synthesis and puppeteering examples of the baselines.

F.4. Limitations

There are multiple scenarios where latent-based video encoding may fail: (1) due to limitations inherited from StyleGAN2 (*e.g.*, fixed resolution, entanglements, alignment requirements, texture sticking, *etc.*), (2) during preprocessing if the face is misaligned with respect to StyleGAN2 expectations, (3) extreme facial deformations and profile views, stemming from the low representation in the FFHQ dataset used in training StyleGAN2, (4) possible identity drift in editing StyleFlow or StyleSpace, (5) wearables such as eyeglasses can be challenging in some cases due to remaining latent space entanglement, (6) both latent space inversion and editing are sensitive to occlusions.

G. Potential Negative Societal Impact

Since the proposed pipeline successfully captures the fine, detailed, and expressive facial attributes, it improves the realism of face re-enactment. Thus, our model could be misused to create re-enactments with ill-intent (*e.g.*, defamation) and we strongly oppose such malicious use. The research on detection of DeepFakes have progressively advanced as well [4, 7, 16, 19], and the data from our model could be used to improve such methods, thus reducing the potential negative societal impact.

References

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[2] M.J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision 25, 23–48*, 1997.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CVPR*, 2019.

[4] Manjary P Gangan, K Anoop, and VL Lajish. Distinguishing natural and computer generated images using multi-colorspace fused efficientnet. *Journal of Information Security and Applications*, 68:103261, 2022.

[5] Shanyan Guan, Ying Tai, Bingbing Ni, Feida Zhu, Feiyue Huang, and Xiaokang Yang. Collaborative learning for faster stylegan embedding. *arXiv preprint arXiv:2007.01758*, 2020.

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[7] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. Countering malicious deepfakes: Survey,

Figure 8. **Example face deformations resulting from manipulation of each StyleSpace index, $(l, c) \in \mathcal{V}$ in the negative (-) and positive (+) directions**. It could be seen that the identity is preserved across all manipulation examples

battleground, and horizon. *International Journal of Computer Vision*, pages 1–57, 2022.

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[11] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[13] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021.

[14] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[15] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[16] Xin Wang, Hui Guo, Shu Hu, Ming-Ching Chang, and Siwei Lyu. Gan-generated faces detection: A survey and new perspectives. *arXiv preprint arXiv:2202.07145*, 2022.

[17] Jianjin Xu, Zheyang Xiong, and Xiaolin Hu. Frame difference-based temporal loss for video stylization. *arXiv preprint arXiv:2102.05822*, 2021.

[18] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[19] Mingxu Zhang, Hongxia Wang, Peisong He, Asad Malik, and Hanqing Liu. Exposing unseen gan-generated image using unsupervised domain adaptation. *Knowledge-Based Systems*, 257:109905, 2022.

[20] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.