

# Auto-CARD: Efficient and Robust Codec Avatar Driving for Real-time Mobile Telepresence

Yonggan Fu<sup>1\*</sup>, Yuecheng Li<sup>2</sup>, Chenghui Li<sup>2</sup>, Jason Saragih<sup>2</sup>,  
Peizhao Zhang<sup>2</sup>, Xiaoliang Dai<sup>2</sup>, Yingyan (Celine) Lin<sup>1</sup>  
<sup>1</sup>Georgia Institute of Technology <sup>2</sup>Meta

{yfu314, celine.lin}@gatech.edu {yuecheng.li, leo.li, jsaragih, stzpz, xiaoliangdai}@meta.com

## Abstract

*Real-time and robust photorealistic avatars for telepresence in AR/VR have been highly desired for enabling immersive photorealistic telepresence. However, there still exists one key bottleneck: the considerable computational expense needed to accurately infer facial expressions captured from headset-mounted cameras with a quality level that can match the realism of the avatar’s human appearance. To this end, we propose a framework called Auto-CARD, which **for the first time** enables real-time and robust driving of Codec Avatars when exclusively using merely on-device computing resources. This is achieved by minimizing two sources of redundancy. First, we develop a dedicated **neural architecture search technique** called AVE-NAS for **avatar encoding** in AR/VR, which explicitly boosts both the searched architectures’ robustness in the presence of extreme facial expressions and hardware friendliness on fast evolving AR/VR headsets. Second, we leverage the temporal redundancy in consecutively captured images during continuous rendering and develop a mechanism dubbed LATEX to skip the computation of redundant frames. Specifically, we first identify an opportunity from the linearity of the latent space derived by the avatar decoder and then propose to perform **adaptive latent extrapolation** for redundant frames. For evaluation, we demonstrate the efficacy of our Auto-CARD framework in real-time Codec Avatar driving settings, where we achieve a  $5.05\times$  speed-up on Meta Quest 2 while maintaining a comparable or even better animation quality than state-of-the-art avatar encoder designs.*

## 1. Introduction

Enabling immersive real-time experiences has been the key factor in driving the advances of Augmented- and Virtual-Reality (AR/VR) platforms in recent years. Photorealistic telepresence [29, 31, 37, 46] is emerging as a technology for enabling remote interactions in AR/VR that aims

to impart a compelling sense of co-location among participants in a shared virtual space. One state-of-the-art (SOTA) approach, coined Codec Avatars [29], is comprised of two components: (1) an encoder, which estimates a participant’s behavior from sensors mounted on an AR/VR headset, and (2) a decoder, which re-renders the aforementioned behavior to the other parties’ headset display using an avatar representation. Both the SOTA encoder and decoder designs have leveraged the expressive power of deep neural networks (DNNs) to enable the precise estimation of human behaviors as well as the high fidelity of rendering, which are critical for immersive photorealistic telepresence.

Despite its big promise, one of the main challenges posed by photorealistic telepresence is the competing requirements between ergonomics and computing resources. On the one hand, power, form factor, and other comfort factors strictly limit the available computing resources on an AR/VR headset device. On the other hand, the DNNs used in SOTA Codec Avatars are computationally expensive and require continuous execution during a telepresence call. It is worth noting that the limited computing resource on an AR/VR device must additionally be shared with other core AR/VR workloads, such as the SLAM-based tracking service, controller tracking, hand tracking, and environment rendering. Therefore, it is highly desirable and imperative to minimize the computation overhead and resource utilization of Codec Avatars, while not hurting their precise estimation of human behaviors and rendering fidelity. This has become a bottleneck limiting their practical and broad adoption.

To close the above gap towards real-time Codec Avatars on AR/VR devices, existing work has focused on reducing the computational cost of the decoder. For example, PiCA [31] leverages the compute characteristics of modern DSP processors to enable simultaneously rendering up to five avatars on a Meta Quest 2 headset [31]. On the other hand, efficient encoder designs that can fit the AR/VR computing envelope have been less explored, with most existing works assuming off-device computing scenarios. Specifically, SOTA methods for the encoder such as [37, 46] are prohibitively heavy with  $\sim 3$  Giga-floating-point-operations

\*Work done during an internship at Meta.

(GFLOPs) for encoding merely from one image, which is too costly to be continuously executed on SOTA AR/VR headsets. Although cloud-based solutions have been explored as an alternative for other AR/VR use cases, on-device encoder processing for Codec Avatars is particularly desired for telepresence applications as a way to better protect the privacy and overcome internet bandwidth limitations.

In this work, we aim to enable real-time encoder inference for Codec Avatars on AR/VR devices. Specifically, the encoder takes in image data captured from headset-mounted cameras (HMC) and outputs facial expression codes for a Variational Auto-Encoder (VAE) [22], which is used as a decoder following prior works [29, 38]. This target problem is particularly challenging due to two reasons. First, naively reducing the encoder capacity, e.g., by compressing the encoder models to have fewer channels and/or shallower layers, typically results in accuracy degradation, especially for extreme expressions at the tail ends of the data distribution which are often precisely the expressions that contain the most informative social signal. Second, since hardware backends are still nascent for AR/VR use cases, heuristics for hardware-specific optimization may quickly become obsolete. For example, the Qualcomm Snapdragon 865 system-on-a-chip (SoC) [1] on Meta Quest 2 headsets and customized accelerators [40] exhibit different latency/energy constraints. As such, it is important for our developed techniques to be able to automatically adapt to different hardware backends for ensuring their practical and wide adoption, instead of relying on manual optimization strategies that require costly laboring efforts.

To tackle the aforementioned challenges, we develop a framework, dubbed Auto-CARD, for enabling efficient and robust real-time Codec Avatar driving. Auto-CARD *automatically* minimizes two sources of redundancy in the encoding process of SOTA solutions: architectural and temporal redundancy. We summarize our contributions as follows:

- Our proposed framework, Auto-CARD, is **the first** method that has enabled real-time and robust driving of Codec Avatars in AR/VR, exclusively using merely on-device computing resources.
- Auto-CARD integrates a neural architecture search technique that is tailored for **avatar encoding** (AVE-NAS), minimizing potential model redundancy while explicitly accounting for the fast-evolving hardware design trends of AR/VR headsets. AVE-NAS comprises three NAS components: (1) a view-decoupled supernet for enabling distributed near-sensor encoding, (2) a hybrid differentiable search scheme for an efficient and effective joint search, and (3) an extreme-expression-aware search objective.
- To further reduce temporal redundancy towards real-time encoders for Codec Avatars on AR/VR devices, Auto-CARD additionally integrates a mechanism,

dubbed LATEX, to skip the computation of redundant frames. Specifically, we first identify an opportunity from the linearity of the latent space determined by the avatar decoder and then propose to perform adaptive **latent extrapolation** for redundant frames.

- Extensive experiments on real-device measurements using AR/VR headsets, i.e., Meta Quest 2 [32], show that our method can achieve a  $5.05\times$  speed-up while maintaining a comparable or even better accuracy than SOTA avatar encoder designs.

## 2. Related Work

**Codec Avatars.** Traditional methods for photorealistic human face modeling [2, 39] rely on accurate but complex 3D reconstruction processes, which are not suitable for real-time applications. To enable photorealistic telepresence, [30] uses a deep appearance model in a data-driven manner, which has been dubbed a Codec Avatar. It adopts a conditional variational auto-encoder [21] to jointly model both the face geometry and appearance, where the encoder encodes the facial behavior into latent codes, which are then decoded back to the facial mesh and view-dependent texture by the decoder. To enhance the rendering quality of gaze and eye contact, which are crucial for immersive face-to-face interactions, [37] explicitly models human eyes' geometry and appearance on top of [30]. Later works further enhance the photorealism [11, 46] of the rendered avatars or enable more diverse functionalities [3, 5, 36, 48]. For example, [46] leverages the constraints imposed by the multi-view geometry to establish precise correspondence between HMC images and avatars and [11] replaces the holistic models with a learned modular representation to enhance the robustness of facial expressions; [5] further extends the model to support novel lighting environments and [36] animates the face model using audio and/or eye tracking.

Despite promising advances in Codec Avatar representation, relatively few studies have been dedicated to on-device generation. PiCA [31] has pioneered efficient decoder design, but it does not address on-device encoding. While PiCA has allowed real-time performance, the strict computational constraints on the device leave little room for encoding and other workloads as decoding five avatars using PiCA will drain out almost all resources on Meta Quest 2 [31]. Our work is the first to directly address the complexity of on-device encoding to enable real-time Codec Avatar driving.

**Neural architecture search.** NAS [51], a sub-field of AutoML [18], automates the design of optimal DNN architectures in a data-driven manner instead of relying on manually hand-crafted heuristics. It has been applied to many different domains [8–10, 12, 17, 23, 27, 28, 42, 44]. Early NAS works target SOTA task accuracy at the cost of prohibitive search time [35, 51, 52] and later innovations adopt weight sharing [4, 6, 16, 28, 33, 49] to greatly reduce the search cost.

Amongst these, differentiable NAS [28] is particularly effective as both the model weights and architectures are updated differentially, greatly boosting search efficiency. In parallel, hardware-aware NAS [7, 13, 17, 42, 44, 45, 47, 50] explicitly accounts for model efficiency during the search process.

Although NAS appears to be a good match for optimizing encoder performance for on-device computing, we find that a straightforward application results in degraded accuracy, especially for extreme and uncommon facial expressions, which are rich in social signals. Additionally, off-the-shelf NAS-optimized architectures from existing works are not optimized for AR/VR hardware. Therefore, we develop a differentiable NAS technique dedicated to the design of Codec Avatar encoders by explicitly incorporating both extreme-expression awareness and hardware awareness.

### 3. Preliminaries about Codec Avatars

#### 3.1. Codec Avatars: Inference

Our work is built on top of [37], a SOTA Codec Avatar model which explicitly models both the geometry and appearance of eyes to achieve immersive eye contact. Specifically, the HMC captured images on AR/VR headsets are first encoded into latent codes  $z$ , gazes  $g$ , and key points  $y$ <sup>1</sup>. Next,  $z$  and  $g$  are transmitted to the receiver end and decoded into face geometry  $G_f$ , view-dependent texture  $T_f$ , and their counterparts for the explicit eyeball model,  $G_e$  and  $T_e$ , using separate decoders  $D_f$  and  $D_e$ :

$$[G_f, T_f] = D_f(z, v, g), [G_e, T_e] = D_e(g, v, e(G_f)), \quad (1)$$

where  $v$  is the view directions and  $e$  extracts the vertex positions of the eyelids based on the face geometry  $G_f$ . The final avatar is rasterized by a differentiable renderer  $R$ , i.e.,  $I = R([G_e, G_f], [T_e, T_f])$ . For more details, we strongly refer the readers to [37].

#### 3.2. Codec Avatars: Training

Simultaneously capturing data for building Codec Avatar encoders and decoders is challenging. This is because a participant’s face is occluded from outside-in cameras while wearing a headset. In addition, the wearing of a headset adds additional effects, such as pressure and restrictions on the hairstyle as well as shading effects, that are not desirable for avatar modeling purposes. As such, existing works [30, 37] first train a face (and eyeball) decoder using a headset-free capture in a multi-view outside-in camera system, and then perform a second capture while wearing the headset. Meanwhile, correspondences between the HMC images and the decoder’s latent space are found by an inverse-rendering process with a domain-transfer component to account for

<sup>1</sup>Although key points are not used by the decoder, they serve as auxiliary supervisions for encoder training and can be used by other AR/VR tasks.

the differences in lighting and spectra between the HMC images and the outside-in camera images used to build the decoder [37]. A real-time encoder can then be trained to map a subset of the HMC images to their corresponding latent codes, gaze, and key points.

## 4. The Proposed Auto-CARD Framework

### 4.1. Framework Overview

The prohibitive computational cost of existing Codec Avatar encoding models is attributed to both the model complexity of per-encoding inference (i.e. model redundancy) and the high encoding frequency for ensuring continuous rendering during remote telepresence (i.e. temporal redundancy). Our method, Auto-CARD, aims to *automatically* minimize these two sources of redundancy while maintaining accuracy via two components: (1) AVE-NAS for discovering an efficient and effective encoder architecture in Sec. 4.2 and (2) LATEX that adaptively skips computation of redundant frames via latent code extrapolation in Sec. 4.3. An overview of Auto-CARD is shown in Fig. 1.

### 4.2. AVE-NAS: Towards Efficient Avatar Encoders

#### 4.2.1 Unique Challenges for Avatar Encoder Search

To ensure search efficiency for enabling fast adaptation to the evolving AR/VR hardware backends, our AVE-NAS adopts a differentiable search scheme. However, directly extending existing differentiable NAS [28] to the avatar encoding process will lead to non-robust encoders, which can achieve a low average error but fail to capture extreme expressions due to the view-collapse issue analyzed in Sec. 4.2.4. Additionally, it may also result in hardware-unfriendly encoders that do not align with the hardware design trends and characteristics of AR/VR headsets, causing undesired high energy consumption. Therefore, we explicitly enforce both robustness awareness to extreme expressions and hardware-friendliness to AR/VR headset hardware in the three components of AVE-NAS as elaborated below.

#### 4.2.2 AVE-NAS: The Search Space

**Considering the trend of distributed near-sensor encoding.** For building the macro-structure of the encoder, one intuitive choice is to transmit different views captured by different HMCs to a central SoC, which will concatenate those views to be jointly processed by the encoder. However, data streaming from peripheral sensors to the central SoC is costly in terms of energy consumption, especially with high frame rates for achieving smooth facial animation [34]. This can not only pose severe challenges for the limited battery life on AR/VR headsets but also continuously occupy the bus resources on the headset during telepresence. As such, the desired future trend is to adopt a distributed near-sensor en-

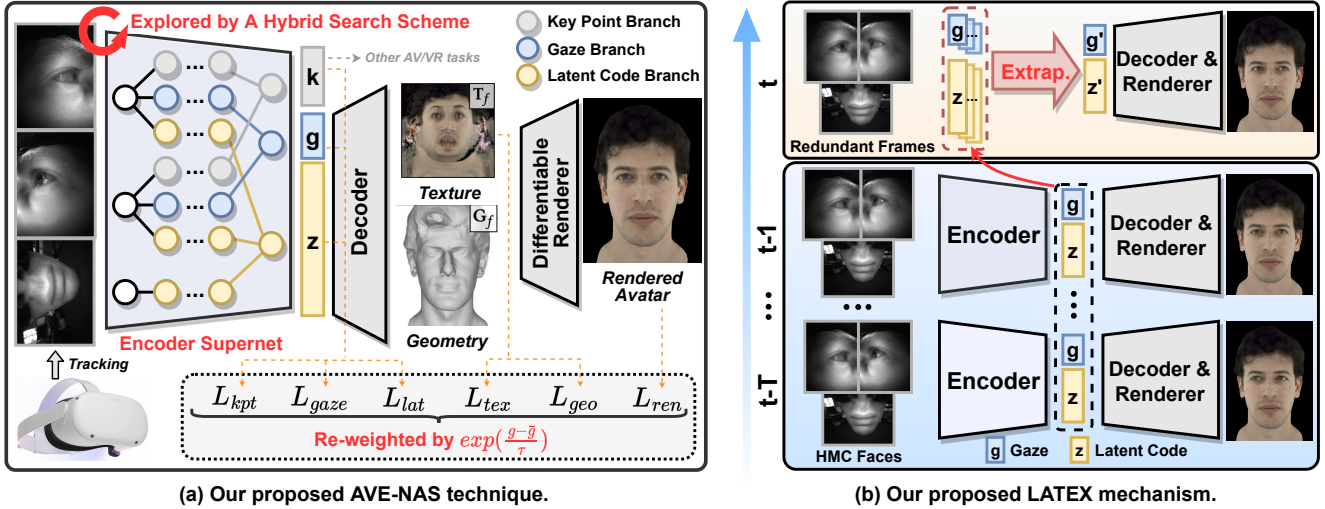


Figure 1. An overview of our Auto-CARD framework, integrating the proposed (a) AVE-NAS and (b) LATEX techniques for minimizing the model and temporal redundancy, respectively.

coding of different captured views [40], which has inspired and motivated AVE-NAS’s supernet design.

**Macro-structure of our supernet.** Our AVE-NAS constructs a view-decoupled supernet that independently processes each of the captured partial-face images. As shown in Fig. 1 (a), we build a three-branch structure for the left and right eyes to generate the gaze, key points, and latent features, respectively. For the mouth view, we adopt only one branch to generate latent features as it is independent of gazes and key points. Finally, the latent features from different views, each of which is a 128-d vector, are concatenated and then regressed into the final latent code. Note that the data movement cost of three 128-d vectors is much smaller than that of transmitting three images (e.g.,  $192 \times 192$  per view in [38]). Considering different cameras here capture different aspects of the facial appearance and motion, the architectures for the aforementioned three views are optimized separately and exhibit different complexity.

**Searchable factors.** To ensure sufficient flexibility of the encoder architecture, the search space spans operator types, depth, width, and input resolution. In particular, our AVE-NAS supports Fused-MBConv [43], single convolution, or skip connections as potential operator types, driven by their high execution efficiency on AR/VR headsets.

### 4.2.3 AVE-NAS: The Search Algorithm

**Challenges of the differentiable joint search for avatar encoders.** Although the operator and channel numbers in our target search space can be differentially searched via the commonly used reparameterization trick [45], the rendering loss is naturally non-differentiable w.r.t. the input resolution. Nevertheless, searchable input resolutions are highly desired for designing avatar encoders, because the structure/texture information captured from different identities often shows diverse complexity which allows using different resolutions for minimizing the overall model complexity. While [45]

achieves differentiable resolution search for classification tasks by inserting paddings inside the input images, such a strategy is not applicable for avatar encoders as it will destroy the structure of the captured human faces. To this end, our design aims to differentially search for input resolutions, together with the operator types and the number of channels, while maintaining the structure information of human faces.

**Our rationale.** To achieve the target design above, we first formulate the joint search process as learning the sampling distributions  $p(\cdot|\theta)$  parameterized by  $\theta \in \{\theta^{op}, \theta^{ch}, \theta^{res}\}$ :

$$\arg \min_{\theta, w} \mathcal{L}(\theta, w) = \mathbb{E}_{p(\alpha|\theta)} [f(\alpha, w)], \quad (2)$$

where  $\alpha \in \{\alpha^{op}, \alpha^{ch}, \alpha^{res}\}$  is the sampled design from  $p(\alpha|\theta)$ ,  $f$  is the objective function, and  $w$  denotes the supernet weights. *The core question* is how to estimate the gradient  $\Delta\theta$  for updating  $\theta$ . Inspired by [15], we find that both reparameterization tricks [21] and policy gradients [41] can produce unbiased gradient estimators for  $\Delta\theta$  in Eq. 2, where the latter does not require  $f$  to be differentiable w.r.t.  $\alpha$  and thus is well-suited for the desired resolution search. Therefore, our AVE-NAS adopts a hybrid differentiable search scheme to integrate both the above estimators.

**Proposed Method.** To implement the aforementioned rationale, for the operator/channel search, we adopt the reparameterization trick from [21] to estimate the gradients  $\Delta\theta^{oc} = \frac{\partial f(\alpha, w)}{\partial \theta^{oc}} = \frac{\partial f(\alpha, w)}{\partial T(\theta^{oc}, \epsilon)} \frac{\partial T(\theta^{oc}, \epsilon)}{\partial \theta^{oc}}$ , where  $\theta^{oc} \in \{\theta^{op}, \theta^{ch}\}$ ,  $T(\theta, \epsilon)$  is a continuous function, and  $\epsilon$  is a random variable. In particular, we formulate the output of each layer as a weighted sum of all candidate choices, e.g., for the operator search, the  $(l+1)$ -th layer  $x_{l+1} = \sum_i T(\theta_{l,i}^{op}, \epsilon) \cdot \alpha_i^{op}(x_l)$  where  $T(\theta_{l,i}^{op}, \epsilon) = \frac{\exp[(\theta_{l,i}^{op} + \epsilon)/\tau]}{\sum_i \exp[(\theta_{l,i}^{op} + \epsilon)/\tau]}$  is the Gumbel Softmax function [19],  $\epsilon$  is sampled from the Gumbel distribution, and  $\alpha_i^{op}$  is the  $i$ -th operator. For the channel search, we follow the channel masking strategy in [45] for search efficiency.

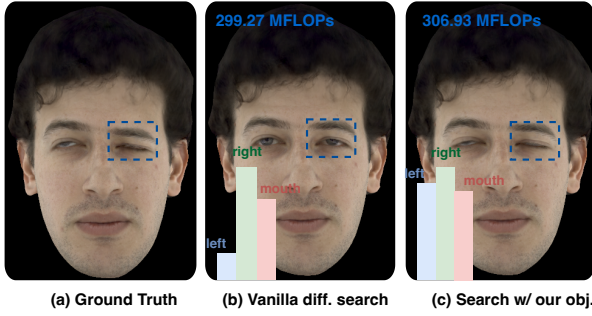


Figure 2. Visualize the rendered avatars by the searched encoders w/o and w/ our proposed objectives, where the normalized FLOPs distributions across three views are annotated.

For the resolution search, we adopt policy gradients [41] to estimate the gradients  $\Delta\theta^{res} = f(\hat{\alpha}, w) \cdot \frac{\partial}{\partial\theta^{res}} \log p(\hat{\alpha}|\theta^{res})$  derived from Eq. 2, where  $\hat{\alpha} \sim p(\alpha|\theta^{res})$  is the sampled resolution. Here  $-f(\hat{\alpha}, w)$  can be viewed as the reward in reinforcement learning [25, 41]. In our design, to make  $p$  differentiable w.r.t.  $\theta^{res}$ , we adopt the Gumbel Softmax function as  $p(\alpha|\theta^{res})$  for sampling  $\hat{\alpha}$ . In addition, to stabilize the search process and reduce the variance in  $\Delta\theta^{res}$ , we sample  $\hat{\alpha}$  once every  $K$  iterations and average the corresponding rewards for updating  $\theta^{res}$ . The searched encoder will be trained from scratch.

#### 4.2.4 AVE-NAS: The Search Objective

**The view-collapse issue.** We empirically find that although NAS-searched encoders can achieve low reconstruction losses on average, they often suffer from notable distortions under extreme expressions, especially those with spatial asymmetry. An example with one eye open and the other closed is shown in Fig. 2 (b), where we can see that the avatar fails to precisely reproduce expression. We discover that while leveraging cross-view correlations (e.g., co-occurring movements between the left and right eyes) to optimize the encoder architecture can lead to reduced model capacity and thus an improved latency, doing so under a latency constraint results in too lightweight structures that fail to capture specific views. The reason is that although those correlations hold for common expressions and thus enable satisfactory average rendering losses, the resulting encoder is non-robust against uncommon and extreme expressions.

**Proposed extreme-expression-aware objective.** To tackle the issue above, we propose an extreme-expression-aware rendering loss for both the encoder search and training to adaptively re-weight the captured faces based on the rareness of the corresponding expressions. Inspired by the focal loss [26], we propose to pay higher attention to the rare facial expressions for boosting the robustness of the searched encoder architecture, whereas the vanilla focal loss in [26] is not applicable for rendering tasks like Codec Avatars. To implement our target design, a proper proxy that can indicate the rareness of the expressions is critical.

As pointed by [37], the rendering quality of gaze and eye contact is the key to achieving immersive face-to-face interactions, while the failure cases with visually unnatural expressions often occur together with inaccurate predictions of the corresponding eye textures. This motivates us to adopt the decoded eye textures as a proxy to indicate the expression rareness. Since the decoded eye textures mainly depend on the predicted gaze as introduced in Sec. 3, we propose to directly adopt the difference between the predicted gaze of the current face  $g$  and the moving averaged gaze  $\bar{g}$  of all previous training faces, the latter of which depicts the most common gaze among all the captured faces, to indicate the rareness of the current facial expression. Formally, we formulate our extreme-expression-aware objective  $\tilde{\mathcal{L}}$  as:

$$\tilde{\mathcal{L}} = \exp\left(\frac{g - \bar{g}}{\tau}\right) \cdot \mathcal{L}, \quad \bar{g} = m \cdot \bar{g} + (1 - m) \cdot g, \quad (3)$$

where we adopt an exponential moving average for performing sample-wise re-weighting. Here,  $\tau$  is a temperature parameter that controls the sharpness and  $m$  is a momentum factor. Note that this same objective is applied in both the encoder search and training stages to avoid view collapse during search and enhance robustness during training.

**Detailed loss design.** In addition to using regressing losses for the latent code, gaze, geometry, texture, and key points in [37], we additionally introduce a rendering loss in the RGB space to search/train the encoders in order to help better refine the high-frequency details of the rendered avatars. This plays a similar role as the perceptual losses in [20]. In particular, the decoded geometry and texture from both the predicted encoding and ground-truth encoding are fed into a differentiable renderer  $R$  to rasterize the avatars into RGB images [37] and for calculating their Mean-Square-Error (MSE) loss. Our final loss function can be formulated as follows, where we omit the coefficients for simplicity:

$$\mathcal{L} = \mathcal{L}_{latent} + \mathcal{L}_{gaze} + \mathcal{L}_{geo} + \mathcal{L}_{tex} + \mathcal{L}_{kpt} + \mathcal{L}_{ren}, \quad (4)$$

where  $\mathcal{L}_{latent} = \|\hat{z} - z\|_2^2$  (similarly for  $\mathcal{L}_{gaze}$  and  $\mathcal{L}_{kpt}$ ),  $\mathcal{L}_{geo} = \|\hat{G}_f - G_f\|_2^2 + \|\hat{G}_e - G_e\|_2^2$  (similarly for  $\mathcal{L}_{tex}$ ), and  $\mathcal{L}_{ren} = \|R([\hat{G}_e, \hat{G}_f], [\hat{T}_e, \hat{T}_f]) - R([G_e, G_f], [T_e, T_f])\|_2^2$ .

**Enforced latency constraint.** During search, we also enforce a latency constraint  $\mathcal{L}_{lat}$ , following [45], based on real-device measurements on AR/VR headsets to control the efficiency of the searched encoders. In particular, we build a measured latency look-up table on Meta Quest 2 for the candidate operators in our search space and use the summed-up latency of the sampled operators as that of the whole network, based on the observation that the runtime of each operator is independent of other operators, following [47].

#### 4.3. LATEX: Leveraging Temporal Redundancy

**Motivation.** During continuous encoding, modeling temporal correlations between latent codes from consecutive

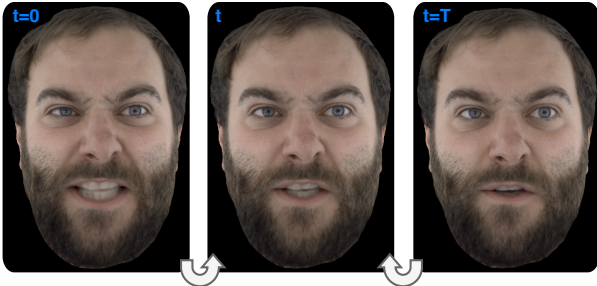


Figure 3. Visualize the rendered avatar (in the middle) decoded from the latent code  $z_t$  based on the interpolation of  $z_0$  and  $z_T$ .

frames can indicate temporal redundancy and thus enable skipping computation for redundant frames. As VAEs [22] are designed for smooth latent spaces, if the latent space of Codec Avatar features good linearity, i.e., linear motions in 3D worlds correspond to linear traversal in the latent space, a simple linear extrapolation can derive the latent code of the current frame from that of previous frames.

#### 4.3.1 Evaluating The Linearity of The Latent Space

We evaluate the linearity of the latent space by linearly interpolating the latent codes of the first and last frames predicted by the encoder, denoted by  $z_0$  and  $z_T$ , respectively, in a batch of size  $T$  to approximate the latent codes of all intermediate frames, i.e.,  $z_t = \frac{T-t}{T} \cdot z_0 + \frac{t}{T} \cdot z_T$ . For temporal windows of size  $T=8$ , we find that this simple approach can achieve a comparable rendering quality and even reduce some jittering effects with an example provided in Fig. 3. This indicates that the latent space determined by the decoder indeed features a decent linearity.

#### 4.3.2 Proposed Adaptive Latent Extrapolation

Considering that the keyframes, which feature a sudden change in facial expressions and/or motions, cannot be linearly extrapolated based on previous frames, an automatic mechanism that can distinguish the key and redundant frames is highly desired. As such, we propose the LATEX technique to adaptively decide whether to derive the latent code via (1) linear extrapolation or (2) encoder inference, using a lightweight early prediction mechanism. In particular, we add an extra prediction head (with  $< 2\%$  FLOPs overhead) to directly estimate the latent code, which can serve as a proxy for identifying key frames. If the difference between the early predicted latent code of the current frame and that of the previous frame is larger than a specified threshold, a complete encoder inference is activated; otherwise, a linear extrapolation is performed to acquire the latent code (i.e.,  $z_t = z_{t-1} + \frac{z_{t-1} - z_{t-T}}{T-1}$ ) based on that of previous  $T$  frames, where the predicted gaze and key points are extrapolated in a similar way. In this way, instantaneous trade-offs between the overall latency and rendering quality can be achieved by varying the threshold.

Table 1. Benchmark the searched encoders with SOTA encoder designs in terms of measured latency on Quest 2/Pixel 3 and rendering MSE across different identities and views.

	Model	EEM	EEM -ch50	EEM -res50	AVE-L (Ours)	AVE-M (Ours)	AVE-S (Ours)
Iden.	MFLOPs	2930.77	765.38	747.44	605.14	306.93	<b>174.75</b>
	Lat. (ms) Quest 2	12.48	10.02	9.40	4.59	3.26	<b>2.47</b>
	Lat. (ms) Pixel 3	483.47	164.27	117.27	70.53	52.61	<b>37.78</b>
S1	Front	8.48	8.54	11.27	<b>6.91</b>	7.46	7.54
	Left	8.11	8.36	10.84	<b>6.80</b>	7.29	7.41
	Right	8.04	8.07	10.65	<b>6.45</b>	7.01	7.03
S2	Front	15.70	16.08	22.77	<b>14.63</b>	15.10	16.21
	Left	14.52	15.08	21.08	<b>13.74</b>	14.56	15.22
	Right	17.51	18.00	24.82	<b>16.17</b>	16.31	17.53
S3	Front	12.03	12.85	15.53	<b>10.91</b>	11.42	12.25
	Left	12.00	12.88	14.96	<b>10.93</b>	11.48	11.99
	Right	12.73	13.62	16.45	<b>11.41</b>	11.83	13.05
S4	Front	17.42	18.71	21.40	<b>15.62</b>	16.33	16.98
	Left	19.12	20.41	23.22	<b>16.80</b>	17.57	18.01
	Right	17.47	18.72	21.27	<b>15.56</b>	16.08	16.94
S5	Front	7.01	7.81	15.95	<b>5.78</b>	5.94	6.06
	Left	7.32	8.05	16.09	<b>6.22</b>	6.33	6.45
	Right	7.10	7.93	15.45	<b>5.92</b>	6.03	6.24
S6	Front	19.52	20.77	25.05	<b>17.34</b>	18.39	19.08
	Left	26.28	26.47	33.76	<b>22.27</b>	24.33	24.34
	Right	15.89	16.73	22.25	<b>14.27</b>	15.41	15.55

## 5. Experimental Results

### 5.1. Experiment Setup

**Dataset.** We adopt the multiview video dataset captured by the face capture system described in [30, 37]. In particular, we adopt  $\sim 12k$  frames (33 camera views per frame) captured by a large multi-camera capture apparatus for training the decoder and  $\sim 42k$  HMC-captured frames (3 camera views per frame) for searching/training/testing the encoder. The HMC captured data of each identity consists of different segments featuring specific expressions, facial motions, and speech, among which we randomly select part of the expression segments as testing data and leave other segments as training data. We evaluate our Auto-CARD framework on 10 identities (4 of which are provided in the appendix) that vary in gender, ethnicity, and age. Encoding estimates are evaluated by decoding into 3 different viewing directions: frontal, left, and right views. The image size of the rendered avatar is set to be  $1024 \times 736$ .

**Devices and measurement settings.** We measure the real-device latency of our searched encoders on two devices with different resource settings, including the SnapDragon 865 SoC [1] on the AR/VR headset Meta Quest 2 [32] and the Google Pixel 3 mobile phone [14]. For SnapDragon 865 SoC on Meta Quest 2, we adopt an in-house compiler to optimize the dataflow for mapping the DNN workloads to the hardware; For Pixel 3, we convert PyTorch models to ONNX, which are then compiled to the TFLite format for execution, following [24].

**Baselines.** We adopt the SOTA encoder design in [37] as

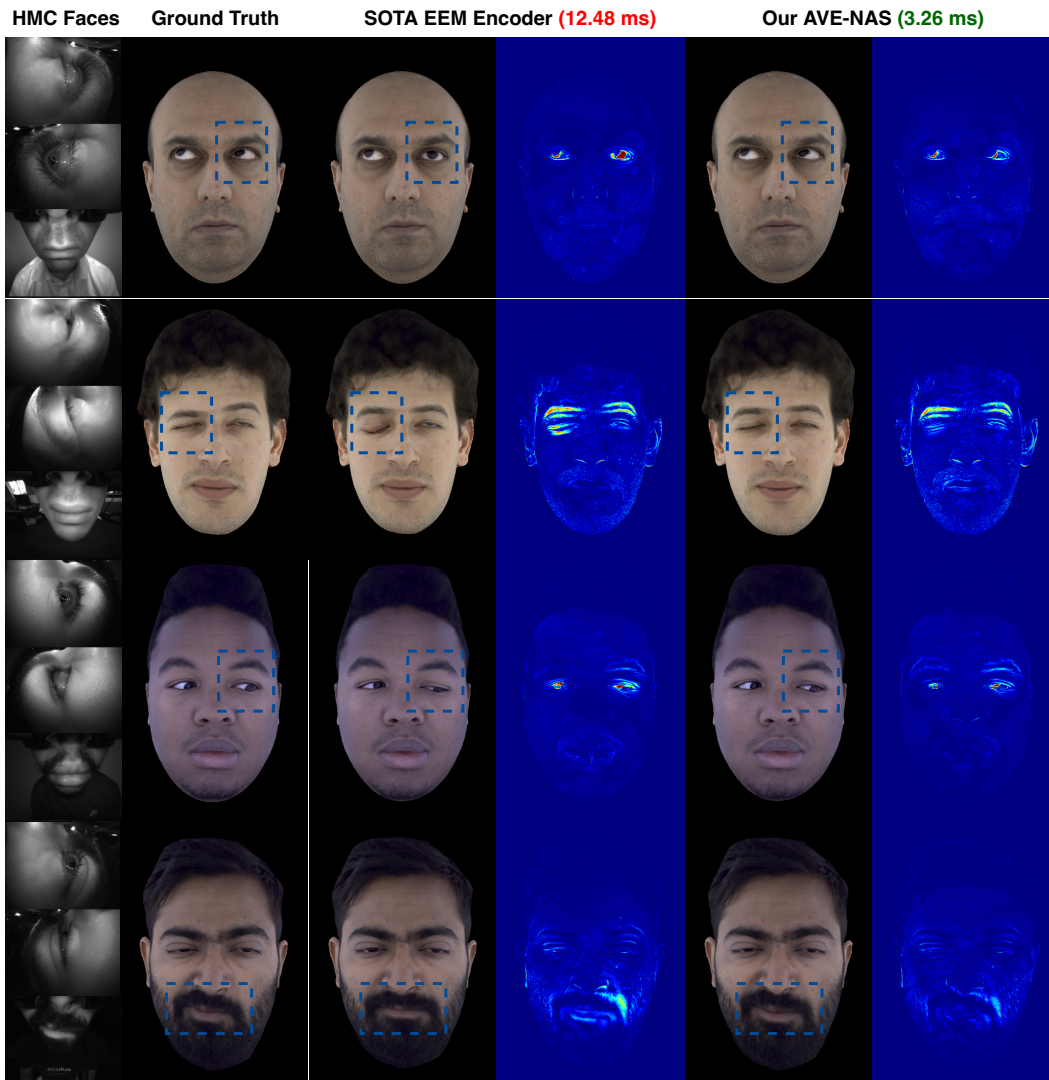


Figure 4. Benchmark the rendering quality achieved by our searched AVE-M against SOTA encoder EEM [38] (zoom-in for better view).

our baseline, which is denoted as EEM. In addition, we also adopt two manually compressed encoder designs, which scale the channel numbers or input resolutions by 0.5 on top of EEM and thus denote EEM-ch50/EEM-res50, as our baselines. We benchmark both quantitative reconstruction quality, in terms of the MSE between the rasterized pixels of the rendered avatars and the ground truth, and qualitative rendering quality, in terms of the visual effects under hard expressions. Detailed encoder structures and hyper-parameters of our Auto-CARD are provided in the appendix.

## 5.2. Evaluating AVE-NAS

**Benchmark with SOTA designs.** By default we perform AVE-NAS on one identity with a rich facial motion under different latency constraints, which are denoted AVE-S, AVE-M, and AVE-L, and then generalize the searched encoder structures to other identities if not specifically stated, aiming at validating the generality of the searched encoders. The details of searched structures are provided in the appendix.

Quantitative comparison: As shown in Tab. 1, we report

the MSE between rendered avatars and the ground truth across 6 identities and 3 view directions. We can observe that: (1) as compared to the SOTA EEM encoder [37], our searched encoders consistently achieve better MSE-efficiency trade-offs across identities based on the measurement on Meta Quest 2, e.g., our searched AVE-L achieves a  $2.72\times$  speed-up over EEM while also reducing the rendering MSE by 1.45 on average and our AVE-S achieves a  $5.05\times$  speed-up with an average MSE reduction of 0.46; (2) our searched encoders achieve notably better compression effectiveness over manually compressed EEM variants, indicating that jointly exploring multiple dimensions of the encoder architecture is crucial for maintaining high rendering quality.

Qualitative comparison: We visualize the rendered avatars across different identities as well as the corresponding difference maps against the ground truth. As shown in Fig. 4, we can observe that our searched encoders can consistently show visually better rendering quality under extreme expressions, i.e., although the SOTA EEM encoder suffers from unnatural facial motions on the eye/mouth re-

gions, our AVE-M can still render photorealistic expressions while achieving a  $3.83\times$  speed-up on Quest 2 headsets.

Table 2. Benchmark the searched encoders by AVE-NAS and its variant VC-NAS in terms of the rendering MSE.

Model	MFLOPs	Lat. (ms)	S1	S2	S3	S4	S5	S6
AVE-M	306.93	3.26	<b>7.46</b>	<b>14.63</b>	<b>10.91</b>	<b>16.33</b>	<b>5.94</b>	<b>18.39</b>
VC-M	299.27	3.07	8.87	17.51	12.26	17.83	7.26	19.47
AVE-S	174.75	2.47	<b>7.54</b>	<b>16.21</b>	<b>12.25</b>	<b>16.98</b>	<b>6.06</b>	<b>19.08</b>
VC-S	200.05	2.53	9.37	17.97	13.17	17.97	7.17	19.66

**Necessity of extreme-expression-aware objectives.** We benchmark the searched encoders by AVE-NAS and its variant without the extreme-expression-aware objective, denoted as VC-NAS since it may suffer from the view-collapse issue analyzed in Sec. 4.2.4, under the same latency constraint. As shown in Tab. 2, AVE-NAS outperforms VC-NAS in terms of rendering MSE across all identities under a comparable latency on Quest 2. Furthermore, we visualize the rendered avatars under expressions with rare gazes in Fig. 5 and observe that the decoded eye textures of VC-NAS may suffer from great distortions under certain expressions and the rendered eyelids/eyeballs notably lose the fidelity caused by the collapsed left- or right-eye branches. In contrast, after introducing extreme-expression awareness, the accuracy is enhanced significantly.



Figure 5. Visualize the rendered expressions of the encoders searched w/o and w/ our proposed objective.

### 5.3. Evaluating LATEX

**Rendering quality.** To validate the rendering quality of LATEX, we plot the MSE-time of the rendered avatars on the test video of one sampled subject w/o and w/ LATEX under different thresholds, which result in different skip ratios (i.e., the ratio of the frames encoded by linear extrapolation to the total frames). As shown in Fig. 6, we observe that our LATEX can skip the encoder inference for 20%~30% frames with a comparable rendering MSE. The detailed LATEX settings are provided in the supplementary material.

**The overall speed-up.** We further show the overall speed-up measured on two devices achieved by combining AVE-NAS and LATEX. In particular, we adopt different skip ratios on top of AVE-S and we report the average MSE in the front view across six identities in Tab. 1 as well as the average

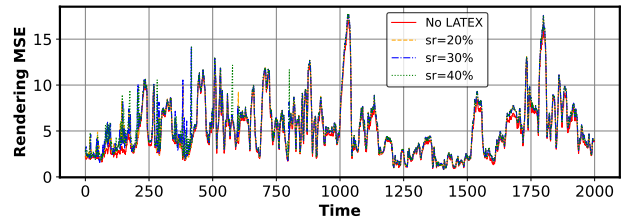


Figure 6. Visualize the rendering MSE evolution w/ and w/o LATEX under different skip ratios (denoted as  $sr$ ).

latency for encoding one frame measured on two devices. We benchmark with the SOTA EEM encoder and a tiny encoder searched by AVE-NAS in Tab. 3 and we observe that: (1) our AVE-NAS can achieve  $6.11\times/15.39\times$  speed-up on Quest 2 and Pixel 3, respectively, with comparable rendering quality (+0.08 MSE), and (2) enabling both AVE-NAS and LATEX can achieve better latency-MSE trade-offs as compared to enabling AVE-NAS only, e.g., a  $1.43\times$  speed-up on average measured on Quest 2 with a 0.48 MSE reduction over AVE-T. This indicates the necessity of LATEX for ensuring the scalability towards ultra-low-latency/resource settings on AR/VR headsets, which is promising for the distributed near-sensor encoding [40] in next-generation AR/VR headsets.

Table 3. The overall speed-up of combining AVE-NAS and LATEX over EEM and AVE-NAS only, where  $sr$  denotes skip ratios.

Model	LATEX	MFLOPs	Quest 2 (ms)	Pixel 3 (ms)	Avg. MSE
EEM	-	2930.77	12.48	483.47	13.68
AVE-T	-	102.60	2.37	32.40	15.72
AVE-S	$sr=0\%$	174.75	2.47	37.48	13.22
	$sr=20\%$		2.06	31.41	13.76
	$sr=30\%$		1.86	28.37	14.42
	$sr=40\%$		1.66	25.33	15.24

## 6. Conclusion

To enable real-time and robust photorealistic avatars on resource-constrained AR/VR headsets, in this work, we proposed Auto-CARD, which for the first time achieves real-time and robust driving of Codec Avatars in AR/VR when exclusively using merely on-device computing resources. In particular, Auto-CARD minimizes both the model and temporal redundancies via the proposed AVE-NAS and LATEX techniques, respectively, while at the same time enhancing the robustness under extreme expressions. Our method can achieve a  $5.05\times$  speed-up on Meta Quest 2 while maintaining comparable animation quality, and our delivered insights can shed light on future innovations in deploying AR/VR-centric computer vision and neural rendering tasks on AR/VR platforms.

## Acknowledgement

The work performed by Yonggan Fu and Yingyan (Celine) Lin is supported by a Meta research award and CoCoSys, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.



## References

- [1] Snapdragon 865 5g mobile platform. <https://www.qualcomm.com/products/application/smartphones/snapdragon-8-series-mobile-platforms/snapdragon-865-5g-mobile-platform>. 2, 6
- [2] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. The digital emily project: photoreal facial modeling and animation. In *Acm siggraph 2009 courses*, pages 1–15. 2009. 2
- [3] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40(4):1–17, 2021. 2
- [4] Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, and Quoc Le. Understanding and simplifying one-shot architecture search. In *International Conference on Machine Learning*, pages 550–559, 2018. 2
- [5] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021. 2
- [6] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*, 2019. 2
- [7] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018. 3
- [8] Daoyuan Chen, Yaliang Li, Minghui Qiu, Zhen Wang, Bofang Li, Bolin Ding, Hongbo Deng, Jun Huang, Wei Lin, and Jingren Zhou. Adabert: Task-adaptive bert compression with differentiable neural architecture search. *arXiv preprint arXiv:2001.04246*, 2020. 2
- [9] Liang-Chieh Chen, Maxwell Collins, Yukun Zhu, George Papandreou, Barret Zoph, Florian Schroff, Hartwig Adam, and Jon Shlens. Searching for efficient multi-scale architectures for dense image prediction. In *Advances in neural information processing systems*, pages 8699–8710, 2018. 2
- [10] Wuyang Chen, Xinyu Gong, Xianming Liu, Qian Zhang, Yuan Li, and Zhangyang Wang. FASTERseg: Searching for faster real-time semantic segmentation. *arXiv preprint arXiv:1912.10917*, 2019. 2
- [11] Hang Chu, Shugao Ma, Fernando De la Torre, Sanjia Fidler, and Yaser Sheikh. Expressive telepresence via modular codec avatars. 2020. 2
- [12] Yonggan Fu, Wuyang Chen, Haotao Wang, Haoran Li, Yingyan Lin, and Zhangyang Wang. Autogan-distiller: Searching to compress generative adversarial networks. *arXiv preprint arXiv:2006.08198*, 2020. 2
- [13] Yonggan Fu, Yongan Zhang, Yang Zhang, David Cox, and Yingyan Lin. Auto-nba: Efficient and effective search over the joint space of networks, bitwidths, and accelerators. In *International Conference on Machine Learning*, pages 3505–3517. PMLR, 2021. 3
- [14] Google LLC. Pixel 3. <https://g.co/kgs/pVRc1Y>, accessed 2020-09-01. 6
- [15] Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. *arXiv preprint arXiv:1711.00123*, 2017. 4
- [16] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. In *European Conference on Computer Vision*, pages 544–560. Springer, 2020. 2
- [17] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1314–1324, 2019. 2, 3
- [18] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated Machine Learning*. Springer, 2019. 2
- [19] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4
- [20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5
- [21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 4
- [22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2, 6
- [23] Royson Lee, Łukasz Dudziak, Mohamed Abdelfattah, Stylianos I Venieris, Hyeji Kim, Hongkai Wen, and Nicholas D Lane. Journey towards tiny perceptual super-resolution. *arXiv preprint arXiv:2007.04356*, 2020. 2
- [24] Chaojian Li, Zhongzhi Yu, Yonggan Fu, Yongan Zhang, Yang Zhao, Haoran You, Qixuan Yu, Yue Wang, and Yingyan Lin. Hw-nas-bench: Hardware-aware neural architecture search benchmark. *arXiv preprint arXiv:2103.10584*, 2021. 6
- [25] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. 5
- [26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5
- [27] 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 Conference on Computer Vision and Pattern Recognition*, pages 82–92, 2019. 2
- [28] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018. 2, 3

- [29] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Trans. Graph.*, 37(4), 2018. 1, 2
- [30] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 2, 3, 6
- [31] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021. 1, 2
- [32] Inc. Meta Platforms. Oculus quest 2, 2021. <https://www.oculus.com/experiences/quest/>, accessed 2021-08-01. 2, 6
- [33] Hieu Pham, Melody Y Guan, Barret Zoph, Quoc V Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. *arXiv preprint arXiv:1802.03268*, 2018. 2
- [34] Reid Pinkham, Tanner Schmidt, and Andrew Berkovich. Algorithm-aware neural network based image compression for high-speed imaging. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 196–199. IEEE, 2020. 3
- [35] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 4780–4789, 2019. 2
- [36] Alexander Richard, Colin Lea, Shugao Ma, Juergen Gall, Fernando de la Torre, and Yaser Sheikh. Audio- and gaze-driven facial animation of codec avatars. In *WACV*, 2021. 2
- [37] Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. The eyes have it: An integrated eye and face model for photorealistic facial animation. *ACM Transactions on Graphics (TOG)*, 39(4):91–1, 2020. 1, 2, 3, 5, 6, 7
- [38] Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. The eyes have it: An integrated eye and face model for photorealistic facial animation. *ACM Trans. Graph.*, 39(4), 2020. 2, 4, 7
- [39] Mike Seymour, Chris Evans, and Kim Libreri. Meet mike: epic avatars. In *ACM SIGGRAPH 2017 VR Village*, pages 1–2. 2017. 2
- [40] H Ekin Sumbul, Tony F Wu, Yuecheng Li, Syed Shakib Sarwar, William Koven, Eli Murphy-Trotzky, Xingxing Cai, Elnaz Ansari, Daniel H Morris, Huichu Liu, et al. System-level design and integration of a prototype ar/vr hardware featuring a custom low-power dnn accelerator chip in 7nm technology for codec avatars. In *2022 IEEE Custom Integrated Circuits Conference (CICC)*, pages 01–08. IEEE, 2022. 2, 4, 8
- [41] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999. 4, 5
- [42] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019. 2, 3
- [43] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021. 4
- [44] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 2, 3
- [45] Alvin Wan, Xiaoliang Dai, Peizhao Zhang, Zijian He, Yuan-dong Tian, Saining Xie, Bichen Wu, Matthew Yu, Tao Xu, Kan Chen, et al. Fbnetv2: Differentiable neural architecture search for spatial and channel dimensions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12965–12974, 2020. 3, 4, 5
- [46] Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. Vr facial animation via multiview image translation. *ACM Trans. Graph.*, 38(4), 2019. 1, 2
- [47] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10734–10742, 2019. 3, 5
- [48] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *arXiv preprint arXiv:2206.15470*, 2022. 2
- [49] Jiahui Yu, Pengchong Jin, Hanxiao Liu, Gabriel Bender, Pieter-Jan Kindermans, Mingxing Tan, Thomas Huang, Xiaodan Song, Ruoming Pang, and Quoc Le. Bignas: Scaling up neural architecture search with big single-stage models. *arXiv preprint arXiv:2003.11142*, 2020. 2
- [50] Yongan Zhang, Haoran You, Yonggan Fu, Tong Geng, Ang Li, and Yingyan Lin. G-cos: Gnn-accelerator co-search towards both better accuracy and efficiency. In *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pages 1–9. IEEE, 2021. 3
- [51] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016. 2
- [52] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018. 2