Supplementary Material:
Self-Supervised Adaptation of High-Fidelity Face Models for Monocular Performance Tracking

Jae Shin Yoon[†] Takaaki Shiratori[‡] Shoou-I Yu[‡] Hyun Soo Park[†]
[†]University of Minnesota [‡]Facebook Reality Labs
{jsyoon, hspark}@umn.edu {tshiratori, shoou-i.yu}@fb.com

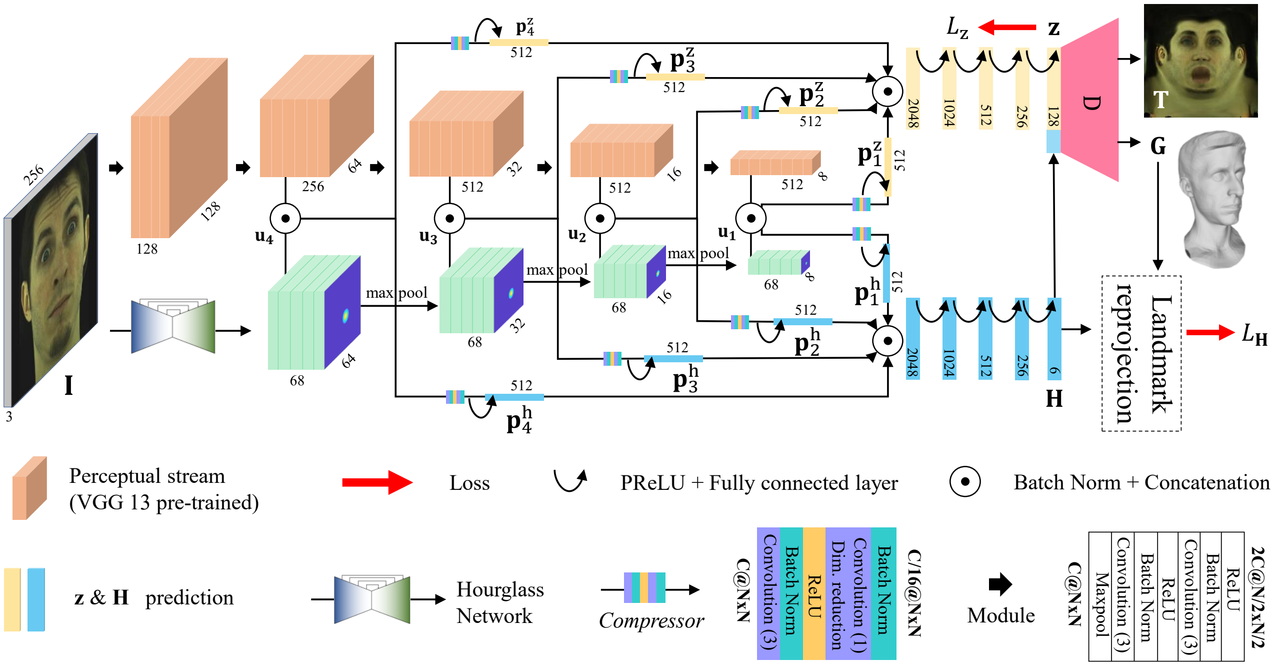

Figure 1: I2ZNet directly regresses the latent facial state codes **z** and headpose **H** from a face image **I**, and the pre-trained decoder D generates full 3D face geometry **G** and high resolution texture **T**.

In the supplementary materials, we provide details on the architecture of I2ZNet in Section A, and the additional ablation studies on I2ZNet will be followed in Section B.

A. I2ZNet

In this section, we detail the architecture of I2ZNet.

Other than only utilizing self-supervised domain adaptation to overcome domain differences, we also explored different networks which could lead to the most domain invariance. Namely, we utilized a combination of a pre-trained VGGNet [5] and HourglassNet [3] to achieve better domain invariance. More details are in Section A.2, and the property of domain invariant features are validated in Section B. Domain specific layers are still necessary to complete our tasks, but thanks to the domain invariant features already extracted by VGGNet and HourglassNet, the domain specific layers can have less parameters thus they are easier to train. We use a combination of deep and shallow features to achieve better performance. More details are in Section A.3.

A.1. Inputs and Outputs

Given a cropped input face image $\mathbf{I} \in \mathbb{R}^{256\times256\times3}$, the I2ZNet directly predicts the low-dimensional facial state codes $\mathbf{z} \in \mathbb{R}^{128}$, and a set of head pose parameters $\mathbf{H} \in \mathbb{R}^6 = \{f,\ r_x,\ r_y,\ r_z,\ t_x,\ t_y\}$, where $\mathbf{f} = \{f\}$, $\mathbf{r} = \{r_x,\ r_y,\ r_z\}$, $\mathbf{t} = \{t_x,\ t_y\}$ are focal length scale, Euler angle, and 2D translation respectively. The pre-trained decoder D decodes $[\mathbf{z}^{\mathsf{T}}, \mathbf{H}^{\mathsf{T}}]$ to generate high fidelity 3D face geometry $\mathbf{G} \in \mathbb{R}^{7306\times3}$ and view dependent texture map $\mathbf{T} \in \mathbb{R}^{1024\times1024\times3}$. Note that, we are using the same decoder with [2], while we replace its encoder network E_X with our I2ZNet.

A.2. Domain Invariant Multi-level Unified Features

Given an input image $\mathbf{I}$, I2ZNet extracts the features from two-stream networks: VGGNet and HourglassNet. VGGNet captures perceptual information such as facial details or shape, while HourglassNet guides "where to look" by providing facial geometry features, e.g. facial landmark heatmaps. We complete the multi-level unified features $\mathbf{u}_l \in \mathbb{R}^{(32*2^l)\times(32*2^l)\times ch_l}$ by concatenating the two-stream features, where $l = \{4, 3, 2, 1\}$ denotes the feature depth-level and the associated channel size is $ch_l \in CH = \{324, 580, 580, 580\}$. Here, we simply max-pool the output from HourglassNet to make the feature size equal to each level of VGG feature. The feature scale inconsistency between two different networks (VGGNet and HourglassNet) is resolved by normalization layer before concatenation. Our multi-level unified features are more domain (color, illumination, or head pose) invariant by learning from domain generalized datasets [1, 4]. Note that, the pre-learned weights on the two-stream networks are fixed in the following training steps such that we prevent I2ZNet from being domain specific.

A.3. Latent Parameter Regression

Inspired by many recent papers [6, 7] which have proposed the use of combination of deep and shallow features to capture semantic-level information and local appearance details at the same time, we concatenate feature vectors from each depth level $\mathbf{p}^z_{4..1}$, $\mathbf{p}^h_{4..1} \in \mathbb{R}^{512}$, which are encoded from $\mathbf{u}_{4..1}$, and they are respectively regressed to $\mathbf{z}$ and $\mathbf{H}$ using several fully connected layers. Here, however, it requires very heavy computational costs for converting three-dimensional features $\mathbf{u}_l$ to single dimensional one $\mathbf{p}^{z,h}_l$ in a fully connected way. Similar to [7], we alleviate this bottleneck by channel-wise feature compression of $\mathbf{u}_l$ to one-sixteenth of its original channel size using two convolutional layers as described as *Compressor* layer in Figure 1.

B. Ablation Studies on I2ZNet

In Section A, we introduced the domain and view invariant property of our network. To verify this, we test I2ZNet on four different scenarios, **View**, **Color**, **Light**, and **Jitter**, where the baseline networks are the same with the ones described in Section 4.2.1.

Table 1: Ablation studies on I2ZNet.

		View	Color	Light	Jitter	Background
VGG Scratch	Geometry	0.607	1.485	1.175	0.983	1.285
	Headpose	-	17.48	6.965	-	15.84
	Texture	-	0.021	0.014	0.016	0.015
VGG	Geometry	1.352	1.258	1.510	1.736	1.076
	Headpose	-	16.61	13.98	-	16.42
	Texture	-	0.020	0.021	0.025	0.016
VGG +Skip	Geometry	0.3967	0.622	0.227	1.331	0.669
	Headpose	-	2.579	0.728	-	8.750
	Texture	-	0.009	0.003	0.018	0.009
VGG +Skip +Key	Geometry	**0.255**	**0.505**	**0.151**	**0.896**	**0.417**
	Headpose	-	**1.676**	**0.684**	-	**8.172**
	Texture	-	**0.007**	**0.002**	**0.012**	**0.006**

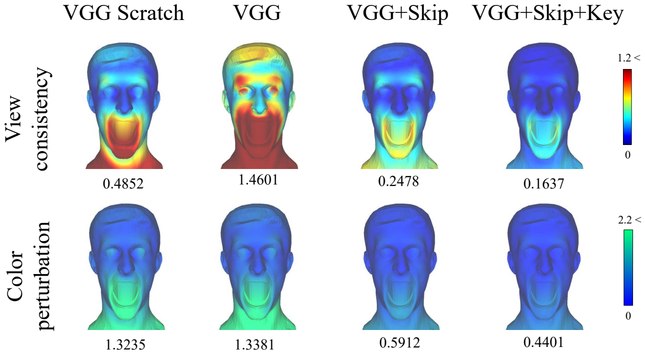

Figure 2: Visualization of the vertex-wise accuracy with a representative subject for ablation studies on view consistency and color sensitivity. The average score is reported for each metric, where the lower score shows the better performance for both scenarios.

View represents the test dataset of multiview videos, where they are accurately synchronized and thus I2ZNet should predict the same facial local deformation to make the facial configuration consistent across the views. To verify this view consistent prediction ability, we pick the most central camera as a ground-truth view and evaluate the performance of other views. We use simple vertex-wise Euclidean distance between the 3D faces predicted from central view and other views meaning that the lower score shows better consistency. The overall performance is summarized in Table 1 and Figure 2, where the proposed network outperforms all other baselines. We can further notice that the combination of skip connection and landmark guidance helps the network to figure out the facial geometry configuration when predicting the facial configuration from different views based on the comparison of **VGG** with **VGG+Skip** and **VGG+Skip+Key**. Note that, when evaluating the view consistency, we remove the texture and head pose from a predicted 3D face because they have view dependent property in our system.

Color, **Light**, **Jitter**, and **Background** represent video sequences which contain synthetic perturbation with random color, gamma, jitters by similarity transformation (scale, rotation, and translation variation), and white dotted background noise. The goal of the test on these sequences is to verify the domain generality. For example, if I2ZNet outputs a completely different 3D facial configuration given a perturbed image comparing to the one before the perturbation, then it implies that the network is overfitted to the training data domain. Therefore, we evaluate the performance of I2ZNet on the sequence after the perturbation in light of the results from the ones before the perturbation. To measure this relative accuracy, we employ three metrics: geometry, texture, and head pose. For geometry and texture, we simply calculate the 3D distance and color difference of the 3D faces. For head pose, we measure the 2D distance between the ground-truth points and the reprojection of the vertices on the 3D face to the input with the predicted head pose. The average scores with respect to the entire test subjects (4 subjects) are reported in Table 1, and the representative subject results are visualized in Figure 2. From the comparison of **VGG Scratch** with **VGG+Skip+Key**, we can notice that the pre-trained nature of the feature extraction parts (VGGNet and HourglassNet) plays a key role to avoid overfitting from a specific domain. Further, the comparison between **VGG+Skip** and **VGG+Skip+Key** implies that the landmark module guides the attention of the network such that it prevents from the network distraction even under the background perturbation.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 2

[2] Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. Deep appearance models for face rendering. *ACM TOG*, 37(4), 2018. 2

[3] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proc. ECCV*, 2016. 1

[4] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image Vision Comput.*, 47:3–18, 2016. 2

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015. 1

[6] Ran Tao, Efstratios Gavves, and Arnold W. M. Smeulders. Siamese instance search for tracking. In *Proc. CVPR*, 2016. 2

[7] Jae Shin Yoon, Francois Rameau, Junsik Kim, Seokju Lee, Seunghak Shin, and In So Kweon. Pixel-level matching for video object segmentation using convolutional neural networks. In *Proc. ICCV*, 2017. 2