Supplementary Material for:
3D human tongue reconstruction from single "in-the-wild" images

In this supplementary material of our paper, we provide linear interpolations between latent features that belong to various random test images. As we can see in Fig. 1, TongueGAN is able to learn meaningful latent feature representations for the unseen space between the source and the target feature vectors. The generated point-clouds are realistic enough with an overall smooth interpolation from the source to the target point-cloud.

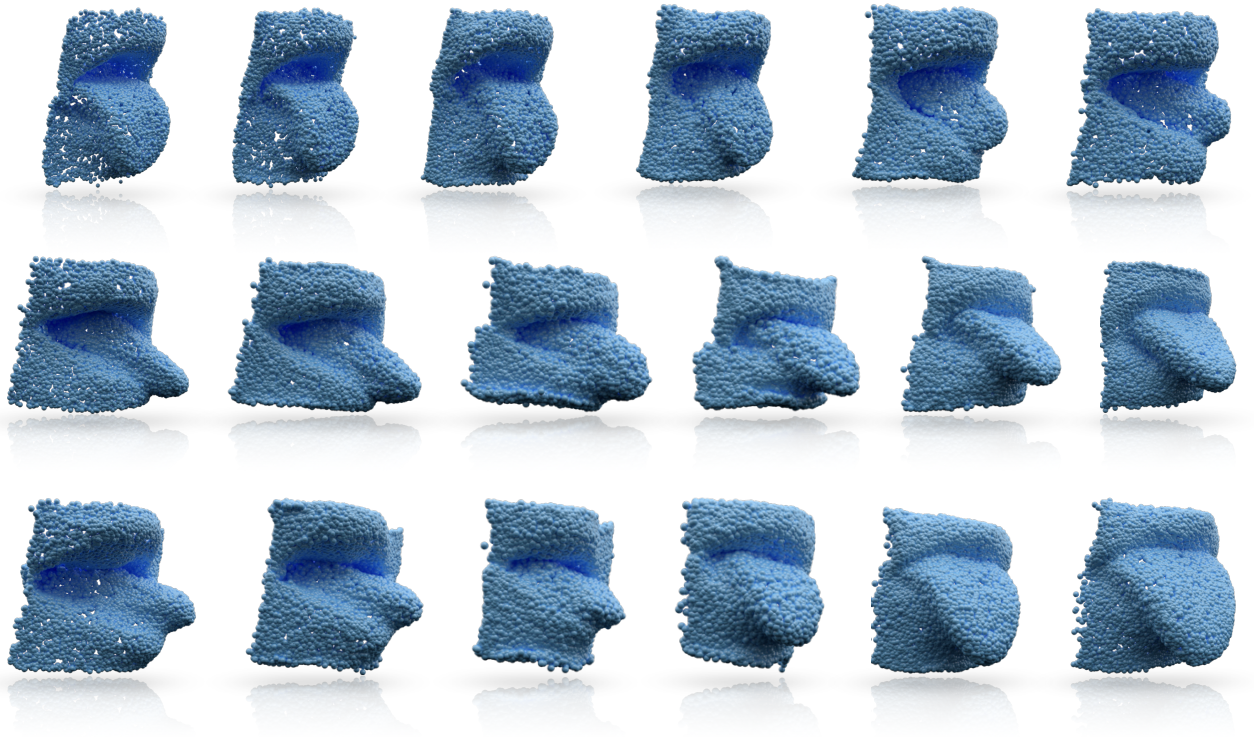

Figure 1. Linear interpolations between the latent features that belong to random test images. The far-left and far-right point-clouds are the source and target point-clouds, respectively.

Moreover, in Fig. 2, we illustrate the 12 mouth landmakrs utilized for the collision detection loss of the main paper.

Finally, in Fig. 3 and Tables 1 and 2 we present the TongueGAN structure and exact architecture. Specifically, in Table 1 we detail the Generator G structure and in Table 2 we detail the Discriminator D structure.

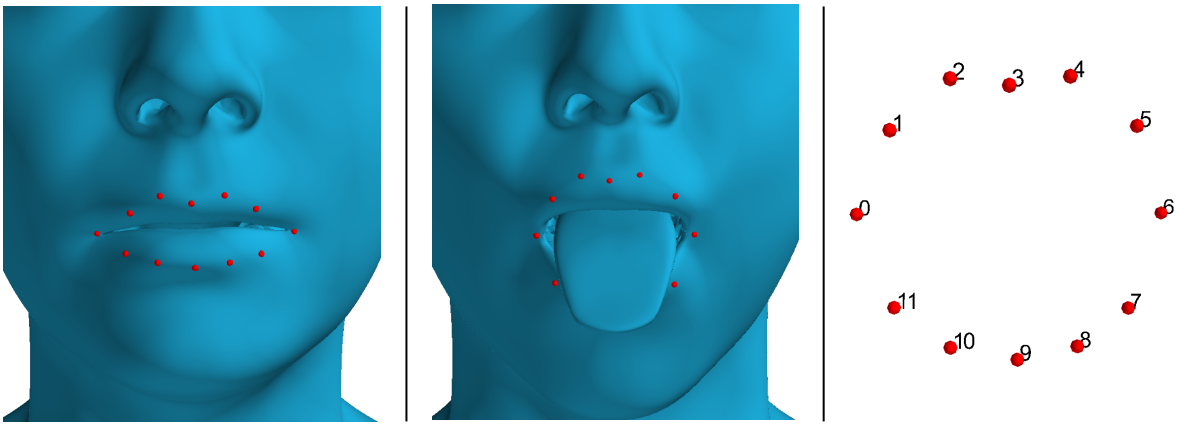

Figure 2. The 12 mouth landmarks of the UHM template which are utilised in equation 1 of the main paper in order to detect collision between the tongue component and the outer oral cavity. These landmarks work as overlapping spheres with radius $r = 1.5cm$.

Table 1. Generator network architecture of TongueGAN.

Part	Layer information
Label MLP	Linear(256, 256), ReLU
Noise MLP	Linear(128, 256), ReLU
Injection MLP Layers	I_1: Linear(512, 512), ReLU
	I_2: Linear(512, 512), ReLU
	I_3: Linear(512, 1024), ReLU
	I_4: Linear(1024, 1024), ReLU
	I_5: Linear(1024, 1024), ReLU
	I_6: Linear(1024, 512), ReLU
	I_7: Linear(512, 512), ReLU
	I_8: Linear(512, 256), ReLU
Main MLP Layers	L_G^1: Linear(512, 512), ReLU
	L_G^2: Linear(512, 512), ReLU
	L_G^3: Linear(512, 1024), ReLU
	L_G^4: Linear(1024, 1024), ReLU
	L_G^5: Linear(1024, 1024), ReLU
	L_G^6: Linear(1024, 512), ReLU
	L_G^7: Linear(512, 512), ReLU
	L_G^8: Linear(512, 256), ReLU
Output Layer	L_G^9: Linear(256, 3)

Table 2. Discriminator network architecture of TongueGAN.

Part	Layer information
Label MLP	Linear(256, 256), ReLU
Point MLP	Linear(3, 256), ReLU
Main MLP Layers	L_D^1: Linear(512, 512), ReLU
	L_D^2: Linear(512, 1024), ReLU
	L_D^3: Linear(1024, 1024), ReLU
	L_D^4: Linear(1024, 1024), ReLU
	L_D^5: Linear(1024, 512), ReLU
	L_D^6: Linear(512, 512), ReLU
	L_D^7: Linear(512, 256), ReLU
	L_D^8: Linear(256, 3)

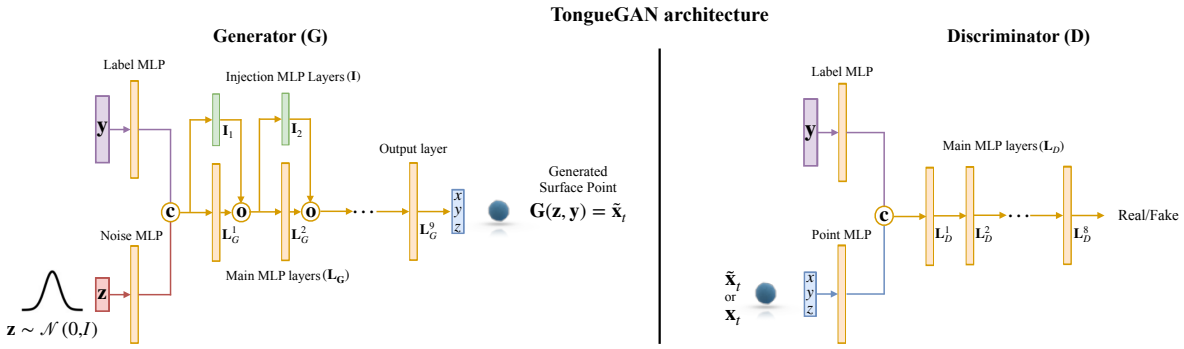

Figure 3. Symbol $\mathbf{c}$ stands for row-wise concatenation along the channel dimension. Symbol $\mathbf{o}$ stands for element-wise (*i.e.*, Hadamard) product. The Generator inputs are a Gaussian noise sample $\mathbf{z}$ and a label $\mathbf{y}$ corresponding to a particular tongue, from which we want to sample a 3D point. The Discriminator input pairs are a label $\mathbf{y}$ which corresponds to a specific tongue and $\mathbf{x}_t$ a *real* 3D point belonging to the aforementioned tongue point-cloud (but sampled as explained in Section 3.3.1 of the main paper) and $G(\mathbf{z}, \mathbf{y}) = \tilde{\mathbf{x}_t}$ a generated point belonging to this tongue. The Discriminator is asked to distinguish the real from the fake point.