Summary of Meshtalk WIP

Anonymous ICCV submission

Paper ID ****

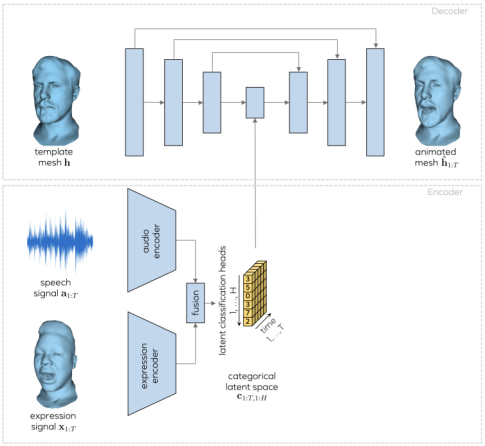

Figure 1. The network diagram[2].

Abstract

The article summerises Meshtalk[2]. The goal of this practice is to get familiar with LaTeX and technical writing.

1. Introuction to Meshtalk

Meshtalk is a generic method for generating full facial mesh animation from speech. It can generate 'lipsync' animation from a single frame of generic human facial mesh, and also can mix in emotional information from mesh animation.

1.1. Network Design

The network resembles Variational Autoencoder with multiple latent space as shown in Figure 1. Target animation estimation $\hat{h}$ is estimated from template mesh$h$ by computing function $\mathcal{D}$.

$$\hat{h}_{1:T} = \mathcal{D}(h, \mathsf{c}_{1:T,1:H}) \tag{1}$$

$c_{1:T}$ is a sequence of encoded latent space derived from audio sequence $\mathsf{a}_{1:T}$ and expression signal mesh $\mathsf{x}_{1:T}$.. c and a is first mapped to $T * H * C$ dimentional latent space, then passed through Gumbel-softmax [1] over every classification head.

$$\mathsf{c}_{1:T,1:H} = [\mathsf{Gumbel}(\mathsf{enc}_{t,h,1:C})]_{1:T,1:H} \tag{2}$$

$$\mathsf{enc}_{1:T,1:H,1:C} = \tilde{\xi}(x_{1:T}, a_{1:T}) \tag{3}$$

1.2. Dataset

1.3. Training

The novel cross-modality loss is defined as

$$\begin{aligned}\mathcal{L}_{\mathsf{x}Mod} = &\sum_{t=1}^{T}\sum_{v=1}^{V}\mathcal{M}_v^{(\mathsf{upper})}(||\hat{h}_{t,v}^{(\mathsf{expr})} - x_{t,v}||) \\ &+ \sum_{t=1}^{T}\sum_{v=1}^{V}\mathcal{M}_v^{(\mathsf{mouth})}(||\hat{h}_{t,v}^{(\mathsf{audio})} - x_{t,v}||)\end{aligned} \tag{4}$$

and loss for eye is defined as following.

$$\mathcal{L}_{\mathsf{eyelid}} = \sum_{t=1}^{T}\sum_{v=1}^{V}\mathcal{M}_v^{(\mathsf{eyelid})}(||\hat{h}_{t,v}^{(\mathsf{expr})} - x_{t,v}||) \tag{5}$$

$\mathcal{M}^{(upper)}$ is a mask that assignes a higer weight to verticies around the mouth, and low weight to other. Correspondingly, $\mathcal{M}^{(mouth)}$ is a mask that assignes a lower weright around the mouth, and vice versa. $\mathcal{M}^{(eyelid)}$ is a specific eye loss, which was crucial.

Final Loss term is defined as $\mathcal{L} = \mathcal{L}_{\mathsf{x}Mod} + \mathcal{L}_{\mathsf{eyelid}}$, which gives both term equal weight.

References

[1] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax, 2017. 1

[2] A. Richard, M. Zollhoefer, Y. Wen, F. de la Torre, and Y. Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement, 2021. 1