

Diversity-Aware Sign Language Production through a Pose Encoding Variational Autoencoder

Mohamed Ilyes Lakhal, Richard Bowden
CVSSP, University of Surrey, Guildford, United Kingdom
{m.lakhal, r.bowden}@surrey.ac.uk



Fig. 1: **Diversity-Aware Sign Language Production.** Given an input image of a signer, we would like to synthesize a novel (unseen) image of another signer given an attribute (*e.g.*, ethnicity) and a corresponding pose. **KEY** – top-row: given a Japanese signer, we produce a Swiss signer with the same pose (Japanese \rightarrow Swiss); bottom-row: Swiss \rightarrow Japanese.

Abstract—This paper addresses the problem of diversity-aware sign language production, where we want to give an image (or sequence) of a signer and produce another image with the same pose but different attributes (*e.g.*, gender, skin color). To this end, we extend the variational inference paradigm to include information about the pose and the conditioning of the attributes. This formulation improves the quality of the synthesised images. The generator framework is presented as a UNet architecture to ensure spatial preservation of the input pose, and we include the visual features from the variational inference to maintain control over appearance and style. We generate each body part with a separate decoder. This architecture allows the generator to deliver better overall results. Experiments on the SMILE II dataset show that the proposed model performs quantitatively better than state-of-the-art baselines regarding diversity, per-pixel image quality, and pose estimation. Quantitatively, it faithfully reproduces non-manual features for signers.

I. INTRODUCTION

In this paper, we present an approach for diversity-aware sign language production, in which we aim to synthesize a signer from a 2D pose sequence and a defined set of attributes (*e.g.*, gender, skin color). To our knowledge, this is the first to tackle this challenging problem. To do so, we introduce a pose-encoding VAE (PE-VAE), a variational inference formulation, to learn the distribution of images of signers

with different attributes. In particular, we explicitly encode the pose information as part of the KL divergence loss. We combine the latent features from variational inference with attribute information. By incorporating the pose into the KL loss, it helps to ensure that the style code does not depend on the pose and learns a better visual feature representation. We propose PENet, a UNet encoder-decoder model with PE-VAE as the visual feature sampler. We generate each body part separately to produce higher visual-quality images. Additionally, we add an edge loss to penalize the synthesis in edge space as we found that this helps the model to produce crisp results.

To summarise, the contribution of this paper is three-fold:

- We propose a new formulation to the Sign Language Production to account for anonymization through pre-defined attributes.
- We extend the Variational inference paradigm to learn pose-agnostic features.
- We implement PE-VAE within the UNet generator where the pose is conditioned with the pose agnostic feature and an attribute to synthesize a signer.

The rest of the paper is presented as follows: Sec. II reviews works related to image synthesis and sign language production. Sec. III introduces our proposed PENet and the

training losses. Sec. IV overviews our experimental protocol and discusses our results against state-of-the-art baselines. Finally, Sec. V summarises the contributions and results of this paper.

II. BACKGROUND

Sign language production. Sign Language Production (SLP) [33] automatically translates a spoken language sequence (Text [43], or Speech [17]) into a sign language sequence. The sequence can either be represented as an intermediate representation (also called Gloss [29]) or directly in a human-readable format (*e.g.*, human pose [14], [36]). Another way to produce Sign Language is to represent the sign through graphical avatars [20]. However, such a solution is less favorable for the deaf community, as small inconsistencies can lead to misleading signs [27]. Therefore, photorealistic SLP [35] was proposed as an alternative to graphical avatars. These models use neural network encoder-decoder architectures to synthesize a sign sequence using identity-specific conditioning [38], [35] or anonymize the signer using variational inference [37], [41].

Neural image synthesis. Neural Radiance Fields (NeRF) [28] predict a scene’s RGB value and volume density based on the position of a 3D point and its viewing direction. Human-NeRF [48], [30], [31] uses a human 3D mesh template (usually SMPL [24]) which is deformed with a given body pose and rendered to the target viewpoint using Volume Rendering [23]. Despite the excellent performance of this class of methods, an incorrect estimation (*e.g.* SMPL) can lead to poor rendering.

Diffusion models [13] are generative models that extend Variational Autoencoders (VAE) [19] to model a complex distribution through a series of Gaussian distributions using a Markovian assumption. We categorise the models into Denoising Diffusion Probabilistic Models [13], where the goal is to synthesize the target in pixel space, and Latent Diffusion Models (LDMs), where the diffusion process takes place in the latent space extracted from a pre-trained model to reduce computational power [34]. In the context of SLP, DiffSLVA [49] takes a video and anonymises it with a text prompt. However, providing the edges from the input video provides enough identity information.

Pose-guided human image synthesis. This problem uses an image and a 2D human pose and synthesises an image with the target pose. The pose could be combined with the image in a single encoder [26] or separate encoders [21]. However, it has been shown that separate encoders are helpful for the overall quality [21].

In terms of architecture, the model is defined as a conditional generative adversarial network (GAN) [7] with the pose and image as input. The input can be represented as 2D pose [26], semantic segmentation [50], or even UV-maps [8]. UNet [22] is largely used with residual blocks [10] as main convolutional modules. Since the input and target poses may differ in terms of scale, orientation, and spatial position, various solutions have been proposed to account for this. Deformable GAN [40] does this by using a mask

extracted from the target pose and applying it to the UNet skip connection in order to force the feature to focus on the spatial positioning of the pose, thus alleviating the problem of pixel misalignment. We can categorise the methods to tackle the input-to-target pose into warping and progressive methods. Warping methods [9], [39] use warping fields to transform (or warp) the feature from the input using the target pose as guidance. Progressive methods [51], [44] attempt to transfer spatial regions (or patches) from the input pose to the target pose.

III. METHOD

This section first provides a problem formulation for diversity-aware Sign Language Production. Then, we relate the use of Variational Auto-Encoder [19] and Vision Transformers [5] to our proposed framework. In particular, we introduce Pose-Encoding Variational Inference (PE-VAE), which explicitly uses the pose and attributes as conditioning in the lower bound computation. The proposed PE-VAE enables us to learn better pose-agnostic features. Finally, we introduce our Pose-Encoding Network (PENet), a UNet encoder-decoder model with multi-head decoders representing each of the head, hand, and torso. PENet takes an input pose and generates an anonymized person based on the conditional attribute.

A. Problem definition

Given a sequence consisting of T RGB frames $\mathcal{X} = \{x_t \in \mathbb{R}^{W \times H \times 3}\}_{t=1}^T$ of a deaf signer, and their corresponding 2D pose $\mathcal{Y} = \{y_t \in \mathbb{R}^{K \times 2}\}_{t=1}^T$ where K is the number of joints, W (resp. H) is the image width (resp. height). We formalise the Diversity-Aware Sign Language Production framework with the help of a generator $\mathbf{G} : \mathbb{R}^{K \times 2} \times \mathbb{R}^d \rightarrow \mathbb{R}^{W \times H \times 3}$, which takes a pose $y \in \mathcal{Y}$ together with an attribute $a \in \mathbb{R}^d$ of feature dimension d , and maps this to the pixel space $x \in \mathcal{X}$, such that: $x \approx \mathbf{G}(x|y, a)$. We do not consider a temporal consistency term since the pose sequence \mathcal{Y} contains the motion information.

B. Preliminaries

Conditional Variational Auto-Encoder. A conditional variational auto-encoder (cVAE) [19] defines a two-stage process: generative and inference. The generative, or encoding, process learns a latent code z from an input x so that the prior distribution is $p_\theta(z|x)$. The latent code z is used to infer the input x during inference or decoding. Because of intractability, we approximate the true posterior with the mapping $q_\phi(z|x)$ also called the variational posterior. By variational inference [19], the evidence lower bound of the log-likelihood term is:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) - \beta D_{KL}(q_\phi(z|x) \| p(z)). \quad (1)$$

The prior $p(z)$ is given as a Gaussian $\mathcal{N}(0, 1)$ and $p_\theta(x|z)$ is defined as the reconstruction term, which is controlled by pixel-wise supervision. The β term controls the estimation of the posterior with the chosen prior [12].

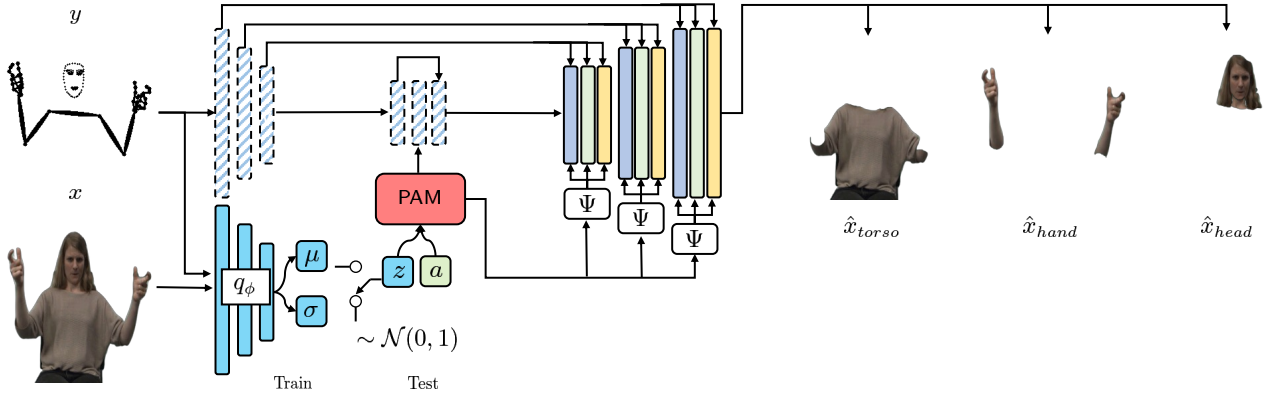


Fig. 2: **PENet**. The network is presented as a conditional VAE-GAN, where the variational parts learn the distribution of visual feature from signers of different attributes (skin tone, ethnicity, gender) through variational inference. The attribute a is presented as a feature vector extracted from a pre-trained CLIP model. The latent code z and a are combined through a MHA module. The pose y is processed through a UNet encoder-decoder network to retain the spatial information of the keypoints, the visual feature z_a guides the synthesis of the person through a mapping Ψ (Eq. 8).

Vision Transformer. The Vision Transformer (ViT) presented in [5] partitions an image x into $N = \lfloor \frac{H}{h} \rfloor \times \lfloor \frac{W}{w} \rfloor$ non-overlapping patches $x_1, \dots, x_N \in \mathbb{R}^{h \times w \times c}$, where w, h are the width and height of the patch respectively. The patches are used as tokens by a linear projection for each x_i as $\mathbf{z}_i \in \mathbb{R}^d$ using a linear operator \mathbf{E} , as $\mathbf{z}_i = \mathbf{E}x_i$.

The tokens are then concatenated into a sequence, together with an additional learnable class token $\mathbf{z}_{cls} \in \mathbb{R}^d$ [4]. A positional embedding $\mathbf{p} \in \mathbb{R}^{(N+1) \times d}$, is also added to the sequence. The tokenization process is thus given as follows:

$$\mathbf{z}^0 = [\mathbf{z}_{cls}, \mathbf{E}x_1, \mathbf{E}x_2, \dots, \mathbf{E}x_N] + \mathbf{p}. \quad (2)$$

The token sequence \mathbf{z} is applied to a transformer encoder with L layers. Each layer, ℓ , is defined as,

$$\mathbf{y}^\ell = \text{Att}(\text{LN}(\mathbf{z}^{\ell-1})) + \mathbf{z}^{\ell-1}, \quad (3)$$

$$\mathbf{z}^\ell = \text{MLP}(\text{LN}(\mathbf{y}^\ell)) + \mathbf{y}^\ell. \quad (4)$$

Where Att is the attention operation using the Query (Q), Key (K), and Value (V) as defined in [45], respectively, LN represents layer norm [1] and MLP is a feed-forward neural network.

C. Pose-Encoding Variational Inference

In the context of diversity-aware SLP, we encourage the generative process to learn the latent code z from the image x , the pose y , and the attribute condition a . We modify the prior distribution as $p_\theta(z|x, y, a)$. Therefore, the pose-conditioned evidence lower bound of the log-likelihood term is updated such that:

$$\log p_\theta(x|y, a) \geq \mathbb{E}_{q_\phi(z|x, y, a)} \log p_\theta(x|z, y, a) - \beta D_{KL}(q_\phi(z|x, y, a) \| p(z)). \quad (5)$$

To implement the conditional posterior q_ϕ , defined in Eq. 5 we use separate encoders for both the pose y and the image x . The feature representation of each input is then combined

using a multi-head attention module as defined in Eq. (2)-(4) such that:

$$\begin{aligned} \mathbf{f}_{xy} &= \text{MHA}(\mathcal{E}_x(x) \oplus \mathcal{E}_y(y)), \\ \mu &= \text{MLP}(\mathbf{f}_{xy}), \sigma = \text{MLP}(\mathbf{f}_{xy}). \end{aligned} \quad (6)$$

Where \mathcal{E}_x (resp. \mathcal{E}_y) is the image (resp. pose) encoder. Compared to the previous state-of-the-art [37], where the conditioning y is omitted during posterior training, we argue that this simplification is suboptimal. Indeed, the latent code z encapsulates the visual features of the signer, and thus, such a representation must be independent of the pose. Therefore, explicitly including the pose y in the conditioning helps to obtain better visual features. Finally, we obtain the latent code by reparameterization as $z = \mu + \sigma\epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$.

We assume that $\mathcal{N}(\mu, \sigma)$ is an uncorrelated multivariate Gaussian random variable of dimension M . Therefore, the VAE loss is given as:

$$L_{\text{VAE}} = \frac{1}{2} \sum_{i=1}^M (1 + \log(\sigma_i^2) - \sigma_i^2 - \mu_i^2). \quad (7)$$

D. Pose Encoding Model (PENet)

Model description. We build our generator \mathbf{G} using a UNet architecture [22]. The reason is that such a class of models retains spatial information from the encoder through skip connections, which is helpful since the network uses the pose y to synthesize an image from the latent code. The latent feature z is combined with the attribute feature $a = \text{CLIP}(\text{attribute})^1 \in \mathbb{R}^{512}$, which is obtained by a pre-trained CLIP model (Fig. 3), so that: $\mathbf{z}_a = \text{MHA}(z, a)$.

The skip connection of a UNet model with L layers is defined as the concatenation between the feature encoder \mathbf{f}_E^i on layer i and the symmetric decoder feature \mathbf{f}_D^{L-i} on the symmetric decoder layer $L - i$.

¹We use a predefined text prompt for each attribute

In the pose-guided human image synthesis problem, the input to the generator is an image with the target pose; in such a case, the encoder contains visual features. In diversity-aware SLP, the encoder contains only coarse key-point features, making it difficult for the network to learn meaningful visual appearance features. A simple solution is to combine the latent code z_a with the last encoding feature in the bottleneck. Such a way might not be efficient since the bottleneck layer loses spatial information. Therefore, we propose to inject z_a into all decoding layers using a mapping Ψ . The skip connection is then defined as follows:

$$\mathbf{f}_{\text{skip}}^{L-i} = \Psi(\mathbf{f}_{\mathcal{E}}^i \oplus \mathbf{f}_{\mathcal{D}}^{L-i}, \mathbf{z}_a). \quad (8)$$

To capture nonmanual features, we opt for a separate head decoder. To retain the consistency of the different body parts, we share the weights across the encoder and the bottleneck layers. The predicted image is therefore obtained as follows:

$$\hat{x} = \hat{x}_{\text{hand}} \cdot \mathbf{m}_{\text{hand}} + \hat{x}_{\text{head}} \cdot \mathbf{m}_{\text{head}} + \hat{x}_{\text{torso}} \cdot \mathbf{m}_{\text{torso}}, \quad (9)$$

where \mathbf{m}_{hand} , \mathbf{m}_{head} , and $\mathbf{m}_{\text{torso}}$ are the hand, head, and torso masks, respectively.

Losses. We use GAN optimisation [7] to train our model. For the generator, we use Perceptual loss [15] to account for low and high frequencies. It uses a VGG₁₉ [42] model and maps the input to a set of L^V higher feature spaces and penalizes the prediction \hat{x} in these spaces. The loss is given as:

$$L_{\text{perc}} = \sum_{l=1}^{L^V} \left\| \text{VGG}_{19}^l(x) - \text{VGG}_{19}^l(\hat{x}) \right\|_2. \quad (10)$$

To further improve for higher frequency details, we use an edge loss that utilizes a combination of Sobel [16], Laplacian [2] filters, and Canny [3] edge detection. The edge loss is then:

$$L_{\text{edge}} = \text{Canny}(x, \hat{x}) + \text{Laplacian}(x, \hat{x}) + \text{Sobel}(x, \hat{x}). \quad (11)$$

In addition, we use a multi-scale discriminator loss [46] with L^D layers, which forces the generator \mathbf{G} to output realistic predictions through a min-max optimization. We also use a conditional discriminator to allow the network to focus on each body part separately. Instead of using a separate discriminator for each body part, we merge them into a single input. The loss is then calculated as follows:

$$L_{\text{feat}} = \sum_{l=1}^{L^D} \left\| \text{tr}^l - \text{pr}^l \right\|, \quad (12)$$

where $\text{tr} = \log \mathbf{D}(y, x_{\text{hand}} \oplus x_{\text{head}} \oplus x_{\text{torso}})$ is the discriminator output from the true sample. The discriminator output from the synthesis of the three parts is $\text{pr} = \log(1 - \mathbf{D}(y, \hat{x}_{\text{hand}} \oplus \hat{x}_{\text{head}} \oplus \hat{x}_{\text{torso}}))$.

We use an attribute loss; the idea is to force the weights of \mathbf{G} to learn discriminative features toward the conditional attribute a . We therefore use a pre-trained ResNet-18 [10]

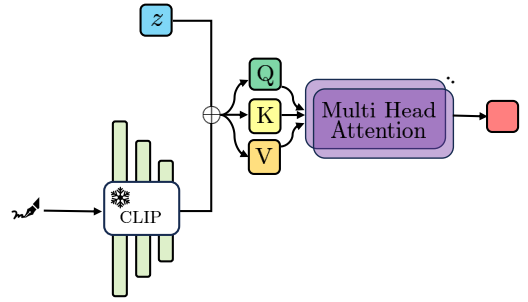


Fig. 3: **Pose Aggregation Module.** Using a text prompt of an attribute (e.g., gender), we extract a feature $a \in \mathbb{R}^{512}$ from a pre-trained CLIP model [32]. This feature is then concatenated with the latent code z and fed into a multi-layer attention module.

where we freeze its weights and add a classification layer. We use cross-entropy to implement the loss L_{attrib} .

In summary, the total loss to train the generator is given as: $L = L_{\text{perc}} + L_{\text{feat}} + \lambda_{\text{edge}} L_{\text{edge}} + \lambda_{\text{attrib}} L_{\text{attrib}} + \beta L_{\text{VAE}}$, where λ_{edge} , λ_{attrib} , β are weighting terms.

IV. EXPERIMENTS

In this section, we evaluate the performance of PENet. We outline the experimental setup and then quantitatively and qualitatively evaluate our model against solid baselines.

A. Experimental setup

Dataset: Due to the difficulty in collecting datasets that are suitable to our need, we choose to use the SMILE [6] sign language dataset as it has a wide range of glosses. Furthermore, we collect additional videos from YouTube to add signers for the ethnicity and skin color attributes. We further process the data containing only atomic glosses (*i.e.* discarding idle frames). We use a weighted data sampler because of the highly imbalanced data distribution. As a data augmentation technique, we add a random rotation of up to 15 pixels, shift, scale, and random horizontal flipping.

Implementation Details: To train PENet, we set $\lambda_{\text{edge}} = 0.01$, $\lambda_{\text{attrib}} = 0.001$, $\beta = 0.001$ and we keep this value fixed throughout the paper unless otherwise specified.

Since no ground truth is available for the attributes (gender, skin tone, and ethnicity), we test the performance of our model in an unconditional setting. Therefore, we drop the attribute and use supervised pixel-to-pixel metrics. In particular, we use Structural Similarity (SSIM), Peak Signal-to-Noise-Ratio (PSNR) [47] as per-pixel measurement, and Fréchet Inception Distance (FID) [11] to assess the diversity of our generative model. To measure the performance of the non-manual features, we use the masked version of each metric for the head, hand, and torso. We also add a pose estimation metric where for each input pose, we sample 5 random samples and run the estimation. We use Mediapipe [25] as a pose estimation model. Furthermore, we add the success rate (denoted by Hit@) when the pose

TABLE I: State-of-the-art comparison with ablation study of our proposed PENet. KEY – Colors: **best**, **second-best**. * We use the original implementation as proposed by the authors.

Model	↑ SSIM			↑ PSNR			↓ FID			
	Head	Hand	Torso	Head	Hand	Torso	Head	Hand	Torso	
Libras [41]	.957 ± .006	.936 ± .020	.873 ± .014	23.32 ± 1.00	21.07 ± 1.58	16.42 ± 1.28	334.12 ± 25.42	306.55 ± 24.42	264.92 ± 23.56	
Anonymsign [37]*	.976 ± .007	.959 ± .015	.944 ± .027	30.58 ± 2.00	26.61 ± 2.41	27.79 ± 3.57	68.11 ± 23.40	127.51 ± 20.82	147.80 ± 25.68	
PENet	w/o L_{edge}	.980 ± .005	.971 ± .012	.949 ± .026	32.55 ± 2.09	29.26 ± 2.31	29.61 ± 3.77	55.09 ± 18.71	88.65 ± 15.64	133.70 ± 23.57
	L_{edge}	.984 ± .005	.970 ± .012	.947 ± .026	33.69 ± 2.39	29.12 ± 2.22	29.33 ± 3.53	46.90 ± 19.07	83.28 ± 16.75	128.06 ± 23.21
	f_{xy} (Conv)	.983 ± .005	.968 ± .014	.948 ± .028	33.35 ± 2.55	28.81 ± 2.54	29.53 ± 3.84	55.34 ± 19.57	95.45 ± 16.59	138.05 ± 27.91
	PE (concat)	.984 ± .005	.970 ± .014	.945 ± .027	33.72 ± 2.88	28.98 ± 2.73	28.06 ± 3.60	51.96 ± 21.46	86.52 ± 19.23	134.08 ± 28.54
	PE (shared)	.984 ± .005	.972 ± .013	.950 ± .028	34.04 ± 2.74	29.77 ± 2.50	29.99 ± 4.11	59.76 ± 27.21	83.70 ± 15.67	129.01 ± 23.99
	PE (separate)	.985 ± .005	.972 ± .012	.951 ± .003	33.98 ± 2.65	29.67 ± 2.46	30.33 ± 3.84	48.37 ± 19.41	77.10 ± 17.67	125.23 ± 23.84
	w/o f_{skip}	.950 ± .007	.912 ± .028	.838 ± .025	24.20 ± 1.27	20.55 ± 1.57	13.77 ± 1.80	224.28 ± 30.42	306.28 ± 30.25	303.46 ± 34.77
	w/o $f_{skip}&\Psi$.947 ± .007	.916 ± .031	.810 ± .020	23.22 ± 1.03	20.33 ± 1.40	12.16 ± 1.96	266.48 ± 28.84	339.12 ± 31.98	307.93 ± 33.22
	Ψ (Conv)	.985 ± .005	.973 ± .012	.950 ± .029	34.21 ± 2.63	30.08 ± 2.24	29.90 ± 4.10	54.56 ± 23.65	82.58 ± 17.20	132.21 ± 26.00
	w/o (m_{hand})	.982 ± .006	.964 ± .012	.947 ± .026	32.18 ± 3.01	27.63 ± 1.76	26.70 ± 3.13	117.88 ± 37.02	147.30 ± 20.89	180.86 ± 34.64
	full model	.984 ± .005	.983 ± .006	.956 ± .027	33.80 ± 2.66	33.38 ± 1.93	31.44 ± 4.51	49.92 ± 15.96	72.07 ± 30.22	125.29 ± 25.37

model successfully estimated the keypoint for the synthesis and the ground truth.

B. Ablation study

Pose representation. The heatmap representation is widely adopted in the literature [26], [35]. Subsequently, it is the default representation for Sign Language. Given a pose $y = \{y_1, \dots, y_K\}$ of K joints, the heatmap \mathcal{H} is defined as:

$$\mathcal{H}_i(y) = \exp\left(-\frac{\|p - y_i\|_2^2}{2\tau^2}\right), \quad (13)$$

where p is the pixel position for an RGB image of size $\mathbb{R}^{3 \times W \times H}$ and $\tau = 6$.

Instead, we propose using RGB-like skeleton images and empirically found that it provides better image quality while considerably reducing the training time by about a half.

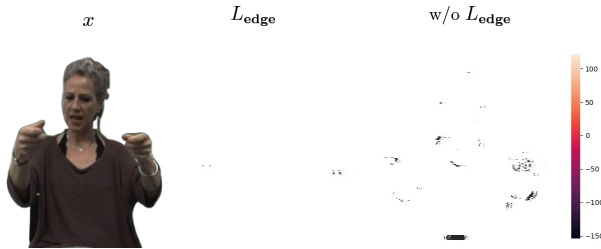


Fig. 4: **Edge loss.** Effect of the edge loss on the synthesised frames. We show the heatmaps of the same frame with and without using the edge loss L_{edge} . Notice the errors when using L_{edge} come mainly from the hand of the clothes *i.e.*, not in the boundaries of the body.

Edge loss. Adding the edge loss L_{edge} contributes to the high-frequency detail and produces crisper results. From Tab. I, we can see that adding the proposed edge loss provides better qualitative results, which are manifested by a lower FID scores against the baseline model. In Fig. 4, we can see that the edge loss has no errors on the body boundaries. Also, we note that the face details are much more visible when using L_{edge} .

Feature representation. To represent f_{xy} , we compare the proposed multi-layer attention module as a feature aggregator in Eq. 6 against convolution. The convolution module noted as f_{xy} (Conv) is defined as a convolution layer with a kernel of size 3, stride of 1, and padding of 1 followed by a

LeakyReLU with a negative slope of 0.2. Results from Tab. I show that the proposed MHA (second row in PENet) produces better results than the convolution module. Furthermore, we also run a similar experiment for the attribute aggregation Eq. 6, and we replace the MHA with a Linear layer to match the dimension. We note similar behaviors as above.

One reason to explain this is the change of each representation. When using CNNs, the kernels are fixed and applied to the input directly, whereas when using MHA, we take into account the input through Query/Key/Value matrices. Hence, since the latent feature z is randomly sampled through a Gaussian at test time, it is more convenient to attend (or weight) the importance of the learned matrices.

Unet representation. The UNet can retain spatial information [22]. In the context of our problem, the network must use skip connections to focus only on painting the semantic body regions. To test this hypothesis, we ran our baseline by removing skip connections; we found that the network could not train properly, and the images consisted of only flickering patches around the poses (Fig. 5). Similar behavior was noticed when training Libras [41].

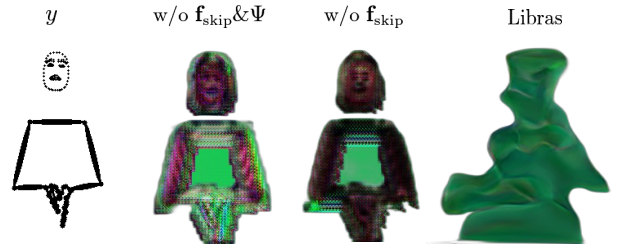


Fig. 5: Example showing the effect of skip connections and injecting the appearance feature z_a using Ψ in the decoder.

Feature aggregator Ψ . We use a StyleGAN [18] feature aggregation to implement Ψ . Additionally, we remove the module in the decoding step noted as Ψ (Conv). From Tab. I, we see that the convolution baseline produces better scores for the head with .985 vs. .984 SSIM and 34.21 vs. 33.80 PSNR. However, our Ψ provides better scores on average, which are translated into the FID scores and better synthesis in general.

Pose-encoding VAE. As explained in Sec. III-C, we explicitly inject the pose information y into the ELBO

TABLE II: Pose estimation results. For each input pose y we sample $z \sim \mathcal{N}(0, 1)$, 5 times and then run the pose estimator.

	AnonySign	PENet	Gain
Head	2.21 ± 3.86	1.13 ± 3.43	1.08 ↓
L_2 R-Hand	21.61 ± 35.72	3.72 ± 13.32	17.89 ↓
L-Hand	12.29 ± 22.21	4.89 ± 16.49	7.40 ↓
Clothes	4.45 ± 6.34	2.81 ± 4.66	1.64 ↓
Hit@	90.87%	91.29%	0.42 ↑
R-Hand	8.28%	38.25%	29.97 ↑
L-Hand	14.44%	22.03%	7.59 ↑
Clothes	99.58%	99.35%	0.23 ↓

calculation. Since there is no ground-truth conditional latent distribution, we relax it to a normal Gaussian distribution $\mathcal{N}(0, 1)$. However, we retain the conditioning on the posterior estimation of q_ϕ . In the following, we empirically verify that retaining the conditioning helps to learn better agnostic visual features z_a .

We propose a late fusion scheme to combine the pose y and the visual feature from the image x . We propose the following ablation to show the effectiveness of such a scheme.

Early-fusion: The pose and the image are concatenated together $y \oplus x$ before being fed to a single encoder that implements q_ϕ .

Shared encoder: The weights of the pose encoder are shared with the pose encoder of the generator \mathbf{G} . The image x has its encoder, and then the feature representation of both inputs is first concatenated and then fed to a multi-layer attention module with 2 layers.

Late-fusion: We follow the same architecture as the shared encoder, except that we do not share the weights of the pose encoder. We note that the three schemes produce similar qualitative results overall. With some differences, as can be seen in Fig. 6. In particular, we found more visible artifacts in the early fusion. This is because x and y are mixed within the same encoder. We see fewer artifacts when we use separate encoders, as with PE(shared) and PE(separate). Using a dedicated pose encoder for the generator \mathbf{G} and q_ϕ provides better qualitative results.

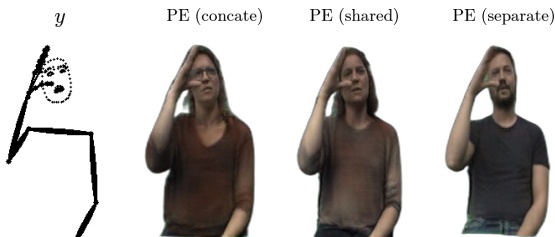


Fig. 6: Results showing various pose encoding VAE schemes as presented in Sec. III-C.

C. State-of-the-art comparison

Our proposed PENet addresses the shortcomings of the two state-of-the-art methods we compare against. The stability of the synthesis as in Libras [41] and the diversity of samples from which AnonySign [37] suffers (Fig. 8).

We address the stability via UNet skip connections as shown in Fig. 5. As for the diversity of the samples, we found that the style loss used in [37] pushes the network to reduce diversity and focus more on the reconstruction loss. Therefore, we decided not to use such a loss in our model.

In Fig. 8, we compare the samples generated by our model with AnonySign. As can be seen, our model has better image diversity and better non-manual features. In particular, our PENet shows distinct facial expressions and accurate hands thanks to our pose-encoding VAE approach.

From Tab. I, we note that the multi-branch approach gives the best results overall. However, we found that adding the hand segmentation helps the network to produce more accurate results. The hand segmentation acts as an explicit mask for each decoder.

The results from Tab. II show that our proposed PENet outperforms AnonySign in skeleton estimation, especially in hand estimation, where our model has 3.72 (resp. 4.89) pixel error for the right (resp. left) hand. AnonySign, on the other hand, has a pixel error of 21.61 (resp. 12.29) pixel error. This is a significant improvement and shows that our model is robust, and we can rely on it for other tasks, such as data augmentation. We also note that our model has a pixel error of 1.13 for the head, which, as mentioned earlier, is suitable for capturing non-manual features. We also found that the Hit@ rate for the hands is lower. This is because the bounding boxes around the hand are small (less than 40×40), and the pose estimator sometimes cannot estimate the hands even for the ground-truth images. Fig. 7 shows the typical behavior of the pose estimation.



Fig. 7: Qualitative comparison of the pose estimation on the synthesised images of signers. The keypoint estimation is highlighted in colors on top of the image. KEY – left: AnonySign; middle: PENet (ours); right: ground-truth.

D. Qualitative analysis

As highlighted earlier, we have no ground-truth samples for the attribute-based synthesis. The main difference with the unconditional synthesis is the additional attribute input. Note that in this case, we replace y in Eq. 6 with $y \oplus a \uparrow$ where $a \uparrow$ is implemented as a 4-layer MLP followed by a 3-layer transpose convolution in order to match with the input



Fig. 8: Method comparison of our proposed PENet against AnonySign [37] using the pose sequence as conditioning.

image dimension of shape $\mathbb{R}^{256 \times 256 \times 3}$. Doing this allows for better modeling of the conditional attribute.

Consistency. We fix the latent feature z_a by randomly sampling from a normal distribution for the input sequence $\mathcal{Y} = \{y_1, \dots, y_T\}$ to evaluate the ability of our model. Fig. 9 shows the results for the gender attribute (female \rightarrow male). Our model can obtain a consistent male appearance while accurately synthesizing the hands and face, encoding the non-manual features. Our experiments show that the encoder-decoder network \mathbf{G} does not need to implement temporal consistency since it is already encoded in the input \mathcal{Y} .

Attribute synthesis It is worth noting that our model can faithfully represent non-manual features of the target attribute. For example, in Fig. 10, our model can synthesize the open mouth (left part) and the head down (top-left). PENet can even generate images with self-occlusion (bottom row). Our model tends to neglect motion blur due to frame-rate sampling. For example, as shown in Fig. 1 (top-left), the synthesised Swiss signer contains no blur, unlike the Japanese signer.

PENet learns a fine-grained representation of body parts without explicit priors. For example, hairstyle is blended with clothing style, while the relevant information from the conditioned attribute remains fixed. It is also crucial for the skin attribute that the skin tone on the arm and face of the synthesised images is uniform.

Challenging cases. Despite the excellent performance of our proposed model, we found some challenging cases. In particular, the influence of the quality of the input pose sequence \mathcal{Y} . Indeed, partial or wrong pose estimation leads to a wrong synthesis. As shown in Fig. 11, the model cannot

cope with partial or missing body parts. Without the head keypoints, the model cannot hallucinate the face (Fig. 11 top). Partial arm estimation also leads to partial synthesis, such as a chopped-off hand (Fig. 11 bottom).

However, we note that the model is robust to the keypoint position. This is a suitable property since correcting the pose sequence \mathcal{Y} can improve synthesis without further model modification. To verify this observation, we perform the following experiment. We change the position of a person’s landmarks and see that the model synthesizes the signer with the updated keypoint position. For example, in Fig. 12, we change the position of the head and can verify that the model generates an image with the changed position of the head.

V. CONCLUSIONS

This paper addresses the problem of diversity-aware Sign Language production. In particular, we address sampling diversity within the same category class and a set of attributes defined in our work, such as gender, skin color, and ethnicity. To achieve this, we propose a GAN-based encoder-decoder framework where we efficiently incorporate the 2D pose conditioning into the variational inference during the training phase. In particular, we use a UNet model as a base generator. The goal is to retain the spatial information from the 2D pose and use visual features from the VAE branch to synthesize an image of a person with the given pose.

The proposed Pose Encoding VAE learns a better agnostic feature representation of signers. Additionally, we can manipulate the attribute to synthesize diverse signers in the same pose. Moreover, our proposed edge loss adds high-frequency details, resulting in better visual quality of images.



Fig. 9: **Synthesis consistency.** By fixing the latent attribute feature z_a , our model can synthesize consistent identities with the skeleton conditioning. This example shows a synthesis from female \rightarrow male.



Fig. 10: **Skin synthesis.** Examples showing the synthesis of skin color signers. KEY – top-row: black \rightarrow white signer; bottom-row: white \rightarrow black signer.

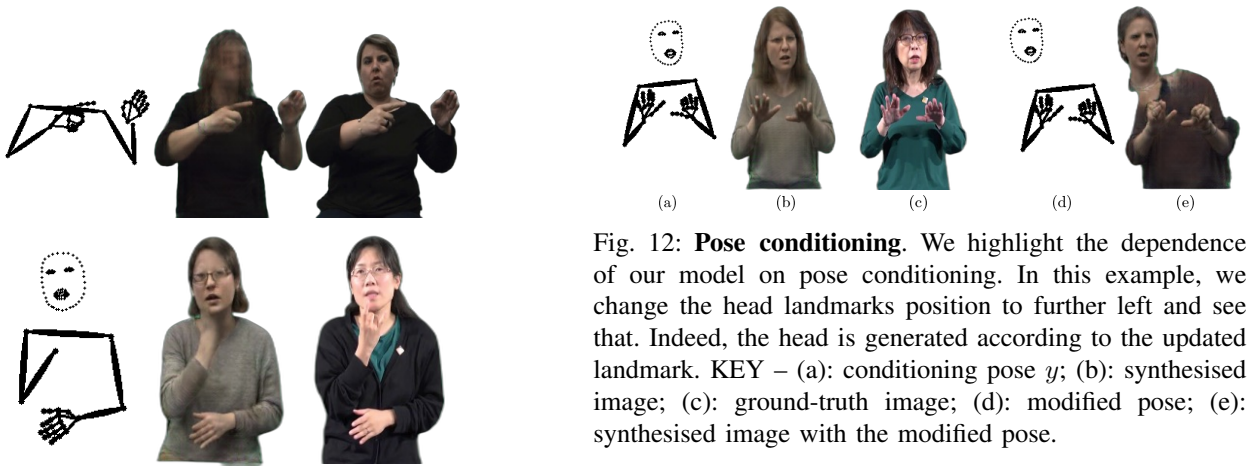


Fig. 12: **Pose conditioning.** We highlight the dependence of our model on pose conditioning. In this example, we change the head landmarks position to further left and see that. Indeed, the head is generated according to the updated landmark. KEY – (a): conditioning pose y ; (b): synthesised image; (c): ground-truth image; (d): modified pose; (e): synthesised image with the modified pose.

Fig. 11: **Failure cases.** We note that partial skeleton estimation leads to failure in the synthesis. KEY – left: input pose; middle: synthesised signer image; right: ground-truth signer image.

We empirically demonstrated that our proposed PE-VAE significantly improves the visual quality of the synthesised images. Extensive experimental evaluations validate the pro-

posed PENet against strong baseline models.

Acknowledgement. This work was supported by the SNSF project ‘SMILE II’ (CRSII5 193686), European Union’s Horizon2020 programme (‘EASIER’ grant agreement 101016982) and the Innosuisse IICT Flagship (PFFS-21-47). This work reflects only the authors view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

- [1] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer Normalization. In *arXiv*, 2016.
- [2] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.
- [3] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, PAMI-8(6):679–698, 1986.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.
- [5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [6] S. Ebling, N. C. Camgöz, P. Boyes Braem, K. Tissi, S. Sidler-Miserez, S. Stoll, S. Hadfield, T. Haug, R. Bowden, S. Tornay, M. Razavi, and M. Magimai-Doss. SMILE Swiss German sign language dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
- [7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. In *Neural Information Processing Systems (NeurIPS)*, 2014.
- [8] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [9] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. VITON: An Image-based Virtual Try-on Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [10] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. In *Neural Information Processing Systems (NeurIPS)*, 2017.
- [12] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.
- [13] J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. In *Neural Information Processing Systems (NeurIPS)*, 2020.
- [14] E. J. Hwang, J.-H. Kim, and J. C. Park. Non-Autoregressive Sign Language Production with Gaussian Space. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021.
- [15] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 694–711, 2016.
- [16] N. Kanopoulos, N. Vasanthavada, and R. Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of Solid-State Circuits*, 23(2):358–367, 1988.
- [17] P. Kapoor, R. Mukhopadhyay, S. B. Hegde, V. Namboodiri, and C. Jawahar. Towards automatic speech to sign language generation. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 6, page 4391 – 4395, 2021.
- [18] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [19] D. P. Kingma and M. Welling. An introduction to variational autoencoders. In *arXiv*, 2019.
- [20] M. Kipp, A. Heloir, and Q. Nguyen. Sign language avatars: Animation and comprehensibility. In *Intelligent Virtual Agents*, pages 113–126, 2011.
- [21] M. Knoche, I. Sarandi, and B. Leibe. Reposing Humans by Warping 3D Features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4499–4504, 2020.
- [22] X. Li, H. Chen, X. Qi, Q. Dou, C. Fu, and P. Heng. H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018.
- [23] B. Lichtenbelt, R. Crane, and S. Naqvi. *Introduction to Volume Rendering*. Prentice-Hall, Inc., USA, 1998.
- [24] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.
- [25] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Ubowaja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann. MediaPipe: A Framework for Building Perception Pipelines. In *arXiv*, 2019.
- [26] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose Guided Person Image Generation. In *Neural Information Processing Systems (NeurIPS)*, 2017.
- [27] J. C. McDonald, R. Wolfe, E. Efthimiou, E. Fontinea, F. Picron, D. Van Landuyt, T. Sioen, A. Braffort, M. Filhol, S. Ebling, T. Hanke, and V. Krausneker. The myth of signing avatars. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, 2021.
- [28] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [29] A. Moryossef, M. Müller, A. Göhring, Z. Jiang, Y. Goldberg, and S. Ebling. An Open-Source Gloss-Based Baseline for Spoken to Signed Language Translation. In *2nd International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, 2023.
- [30] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.
- [31] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *arXiv*, 2021.
- [33] R. Rastgoo, K. Kiani, S. Escalera, and M. Sabokrou. Sign language production: A review. In *arXiv*, 2021.
- [34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [35] B. Saunders, N. C. Camgöz, and R. Bowden. Everybody Sign Now: Translating Spoken Language to Photo Realistic Sign Language Video. In *arXiv*, 2020.
- [36] B. Saunders, N. C. Camgoz, and R. Bowden. Progressive transformers for end-to-end sign language production. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [37] B. Saunders, N. C. Camgoz, and R. Bowden. Anonymsign: Novel Human Appearance Synthesis for Sign Language Video Anonymisation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 2021.
- [38] B. Saunders, N. C. Camgoz, and R. Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5141–5151, June 2022.
- [39] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. Animating Arbitrary Objects via Deep Motion Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [40] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe. Deformable GANs for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [41] W. Silveira, A. Alaniz, M. Hurtado, B. C. Da Silva, and R. De Bem. SynLibras: A Disentangled Deep Generative Model for Brazilian Sign Language Synthesis. In *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, volume 1, pages 210–215, 2022.
- [42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- [43] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden. Text2Sign:

- Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision (IJCV)*, 128(4):891–908, 2020.
- [44] H. Tang, P. H. Torr, and N. Sebe. Multi-channel attention selection gans for guided image-to-image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2022.
- [45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is All you Need. In *Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.
- [46] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [48] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022.
- [49] Z. Xia, C. Neidle, and D. N. Metaxas. DiffSLVA: Harnessing Diffusion Models for Sign Language Video Anonymization. In *arXiv*, 2023.
- [50] C. Xu, Z. Chen, J. Mai, X. Xu, and S. He. Pose- and attribute-consistent person image synthesis. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19, 2023.
- [51] Z. Zhu, T. Huang, M. Xu, B. Shi, W. Cheng, and X. Bai. Progressive and Aligned Pose Attention Transfer for Person Image Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(8):4306–4320, 2022.