2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Let's Play Music: Audio-driven Performance Video Generation

Hao Zhu[1,2], Yi Li[2,3,4], Feixia Zhu[1], Aihua Zheng[1], and Ran He [2,3,4,*]
[1]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation,
School of Computer Science and Technology, Anhui University
[2]Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR), CASIA, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[4]Center for Excellence in Brain Science and Intelligence Technology, CAS, Beijing, China
Email: haozhu96@gmail.com, yi.li@cripac.ia.ac.cn, emmazfx@163.com, ahzheng214@ahu.edu.cn, rhe@nlpr.ia.ac.cn

***Abstract*—We propose a new task named Audio-driven Performance Video Generation (APVG), which aims to synthesize the video of a person playing a certain instrument guided by a given music audio clip. It is a challenging task to generate the high-dimensional temporal consistent videos from low-dimensional audio modality. In this paper, we propose a multi-staged framework to generate realistic and synchronized performance video from given music. Firstly, we provide both global appearance and local spatial information by generating the coarse videos and keypoints of body and hands from a given music respectively. Then, we propose to transform the generated keypoints to heatmap via a differentiable space transformer, since the heatmap provides more spatial information but is harder to generate directly from audio. Finally, we propose a Structured Temporal UNet (STU) to extract both intra-frame structured information and inter-frame temporal consistency. They are obtained via graph-based structure module, and CNN-GRU based high-level temporal module respectively for final video generation. Comprehensive experiments validate the effectiveness of our proposed framework.**

I. Introduction

Given a music audio of a proper instrument, professionals can distinguish which video of a certain person is playing this music, since they have the taught expert knowledge to link the relationship between the music and the corresponding performance actions. Herein, we raise a novel task in this paper: how to generate a performance video of a person playing the given arbitrary music of a specific instrument? We name this task as audio-driven performance video generation (APVG), which has widely potential applications such as concert video generation, instrumental teaching, and VR synthesis. It is a brand-new but challenging task since the extreme hardness to guide the informative motion details such as body and fingers from the heterogeneous low-dimensional audio information.

Prevalent face or body generation models employ keypoints or heatmap to guide the generation [1], [2], [3], [4], [5], [6]. Specifically, the heatmap achieves more impressive performance by offering more spatial information [4], [1], [2], [3]. However, keypoints are much easier to predict from audio [5], [6] since the heatmap is generally sparse and tends to introduce blurry and jittery generation. In order to utilize the advantages of both keypoints and heatmap, we propose to transform keypoints to the heatmap via a differentiable space transformer inspired by [7], then use the informative heatmap to guide the body generation in the performance videos.

Furthermore, conventional works leverage keypoints as a condition to guide the audio-driven video generation [8], [6], which have ignored the rich structure information in the coordinates layout of the keypoints. In order to explore the local structure information during the audio-driven body generation, we further utilize Graph Convolutional Network (GCN) [9], which is one of the prevalent method to encode discrete features with intrinsic structure, to discover the intra-frame structured information from feature blocks.

The key issue of AVPG is to generate temporally smooth performance frames. Conventional video generation schemes either lack of temporal information [3] or leverage computational optical flow to discover the temporal information [10]. UNet [11], which utilizes skip-connections to pass the features to the corresponded decoder layer, has been noted as a prevalent architecture with promising performance in image-to-image tasks [12], [13]. However, the conventional UNet cannot capture the temporal information, which is crucial in video generation. Recently, GRU (Gated Recurrent Unit) [14] has been drawn increasing attention in computer vision tasks due to its ability of providing long term memory of previous frames. Therefore, we propose to concatenate UNet in adjacent frames by propagating high-level feature of current frame to next frame via CNN-GRU to preserve the inter-frame temporal consistency during generation. Conventional GRU leverages FC layers to capture the temporal information while destroying the spatial information in original image space. CNN-GRU replaces the FC layer by Conv layer to preserve the spatial information in high-level features and achieves better performance in practice [6].

Based on above discussion, we propose a multi-stage approach to capture both intra-frame spatial structure and inter-frame temporal consistency for audio-driven performance video generation. The overall architecture and the pipeline of our method is illustrated in Fig. 1. To the best of our

* corresponding author

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412698

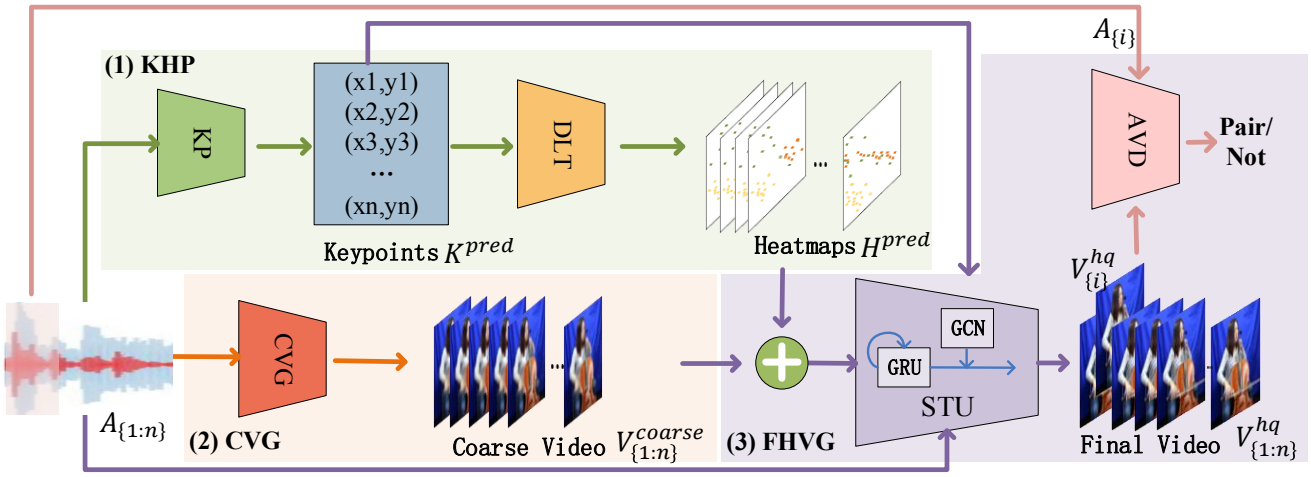


Fig. 1: The pipeline of our proposed model. It contains three main steps: (1) Keypoint and Heatmap Prediction (KHP) which predicts the keypoints from the given music audio clips via Keypoints Predictor (KP), and then transforms the predicted keypoints into corresponding heatmap via Differentiable Landmark Transformer (DLT). (2) Coarse Video Generator (CVG) which generates the coarse video from given audio for further refinements. (3) Final Performance Video Generation (FPVG), which integrates the graph represented intra-frame structure information from predicted keypoints via GCN module and temporal information via CNN-GRU module. The representations concatenated by the generated coarse video and the predicted heatmap, the given audio via the proposed Structured Temporal UNet (STU). We finally feed the pair of each generated video frame and corresponding audio segment into the Audio Video Discriminator for judgement.

knowledge, this is the first work exploring the audio-driven performance video generation (APVG) task. The main contributions of this work can be summarized as:

- We propose an effective multi-stage adversarial generation model to achieve the APVG task, which casts a new challenging problem for audio-visual computation and provides a baseline framework for related researches and potential applications.
- We propose to transform the predicted keypoints to corresponding heatmap by utilizing a differentiable landmark transformer (DLT) to provide more precise local spatial information, followed by the concatenation with the coarse video generated by the given music clips, to provide global appearance information for APVG.
- We propose an Structured Temporal UNet (STU) for the high-quality performance video generation in APVG, which can simultaneously capture the intra-frame structure information via graph-based representation on the predicted keypoints and inter-frame temporal consistency via CNN-GRU connected UNet.

II. Related Work

With the development of Generative Adversarial Networks (GAN) [15], many works leverage this idea to generate images/videos [13], [10] or to assist other tasks [16], [17]. Audio-visual generation [18], as one of the generative tasks, consists of audio guided visual generation [19], [20], [21], [22] and visual guided audio generation [23], [24]. Although APVG is a new task in audio-visual generation, there are some similar tasks such as: music-driven pose motion synthesis, talking face generation and human pose transfer.

A. Music-driven Pose Motion Synthesis

Given the audio music, music-driven pose motion synthesis aims to predict a sequential structure of the body followed rendering or avatar animation to produce the final motion videos. [25] explored the relationship between the music and motion by training a music-motion matching quality rating function. [26] proposed a real-time GrooveNet based on Conditional Restricted Boltzmann Machines (FCRBM) and Recurrent Neural Networks (RNN) to generate dance movement from music. [27] proposed to leverage an auto-regressive encoder-decoder network to generate choreography system from music. [28] proposed to generate keypoints of the body from audio, followed by the avatar animation. Recently, Zhuang et al. [29] leveraged global and local feature to shift the WaveNet [30] from speech generation to the pose motion synthesis. Lee et al. [31] decompose a dance into dance units, and proposed a network to learn how to reorganize these units via given music. However, these methods mainly synthesize the keypoints or skeletons to describe the body motion then generate the motion video by the renderer, while our APVG task directly generates the body motion videos from music.

B. Talking Face Generation.

Given a audio clip, talking face generation aims to synthesize a realistic talking face video with lip synchronization

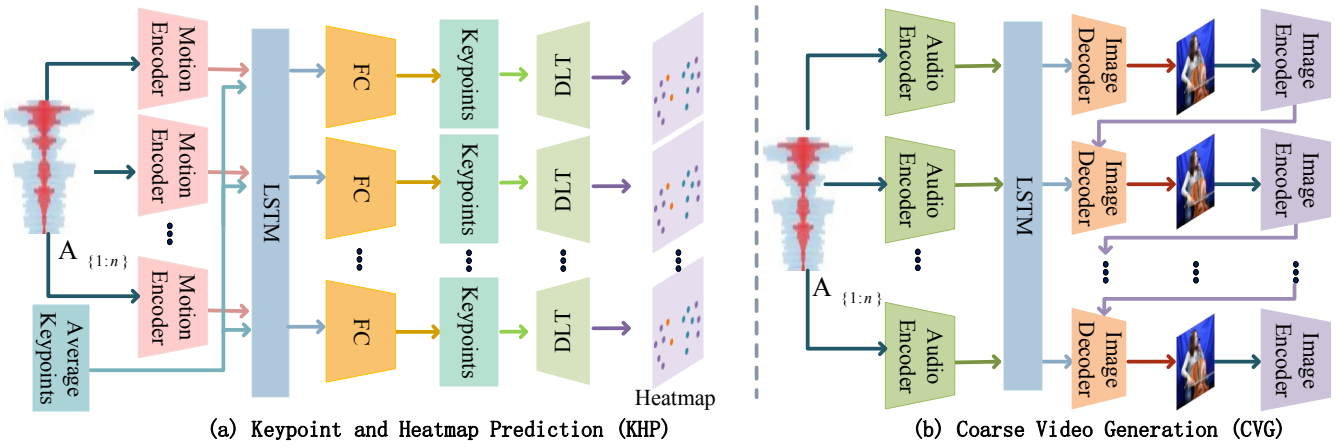


Fig. 2: The pipeline of our proposed (a) Keypoint and Heatmap Prediction (KHP), and (b) Coarse Video Generation (CVG).

of facial motion over the entire video speech. Earlier works synthesized talking face for a specific person [8], [32], while most recent methods focus on the synthesis for arbitrary identity [20], [21], [6]. [20] proposed to disentangle the audio-visual representation into word-related and identity-related representation. [21] introduced mutual information approximation to capture high-level coherence between audio and visual modalities. [6] transferred audio to facial landmarks and then generating attention and motion masks on the landmarks for final video frames. However, the task of synthesizing the global body motion, together local finger motion from the given audio, is more challenging and complex.

C. Human Pose Transfer

Human pose transfer aims to generate the image of a person in arbitrary poses. This task was first proposed by Ma et al. [4] which leveraged the coarse-to-fine scheme to synthesize the target person from the heatmap obtained from 18 keypoints. Balakrishnan et al. [12] divided pose transfer problem into several sub-tasks and synthesized the target foreground and background separately to adapt the complex background scenes. UNet based architecture is a prevalent approach for pose transfer task, while hard to apply for non-aligned objects. Siarohin et al. [33] introduced deformable skip connections to GAN to handle the non-aligned input and output. Pumarola et al. [34] further proposed a fully unsupervised pose generation scheme by mapping the original pose image back from the generated one via a bidirectional generator. However, they mainly devoted to synthesize the high quality image of a person in different poses while lacked of temporal information.

III. Approaches

Given an audio sequence $A_{\{1:n\}}$ (n denotes the number of the audio clip) which contains the music of a proper instrument, our purpose is to synthesize a high-quality performance video V^{hq}. Our method consists of three parts: Keypoint and Heatmap Predictor (KHP), Coarse Video Generation (CVG), and Final High-quality Video Generation (FHVG) As shown in Fig 1, we shall elaborate each part in this section.

A. Keypoint and Heatmap Prediction: KHP

To take both advantage of keypoints (easy to predict) and heatmap (with more spatial information) during the body motion generation, we propose to first predict the keypoints from the given audio via Keypoint Predictor (KP), and then transform the predicted keypoints into corresponding heatmap via Differentiable Landmark Transformer (DLT) for further video generation. Based on the prior experiments that, people act in different motion templates while playing different instruments, we consider two parts features in keypoints prediction: (1) instrument-related feature, to determine the approximate position of predicted keypoints by feeding the average keypoints from the training set into the keypoints predictor. (2) motion-related feature, to predict more precise positions extracted from Motion Encoder (1D-CNNs) on the current audio clip. We concatenate these two features and feed them to LSTM and FC layer to predict the keypoints.

Then, we transform predicted keypoints K^{pred} to corresponded heatmap H^{pred} via the DLT inspired by [7]. H^{pred} is first filled with the scalar value 0, then calculated with the following equation:

$$H^{pred} = \sum_{i=0}^{W}\sum_{j=0}^{H}\sum_{s=0}^{P}\alpha * max(0, 1 - |H_i^{pred} - K_{s_x}^{pred}|) * max(0, 1 - |H_j^{pred} - K_{s_y}^{pred}|), \quad (1)$$

where W, H, and P denote the width and height of the image and number of keypoints respectively. Furthermore, s_x and s_y denote the x-axis and y-axis coordinate of s-th keypoint and $\alpha = 1$ is an intensity factor. The real heatmap H^{real} can be obtained in the same manner. The full pipeline is illustrated as Fig. 2 (a).

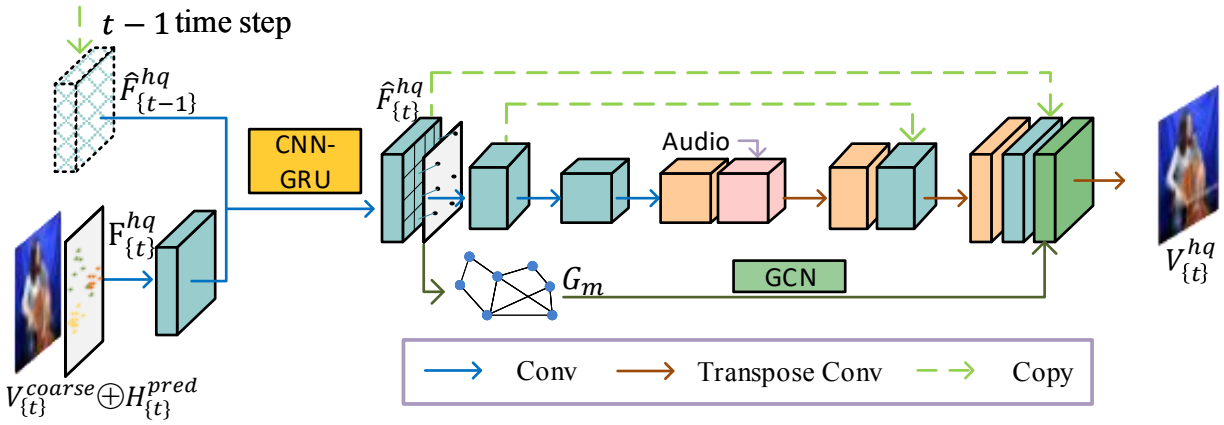


Fig. 3: Illustration of our proposed STU. In the t-th time step, we first extract $F^{hq}_{\{t\}}$ from V^{coarse} and H^{pred}, then fuse F^{t}_{1} and $F^{hq}_{\{t-1\}}$ by CNN-GRU which produces $\hat{F}^{hq}_{\{t\}}$. Second, we use K^{pred} to construct adaptive graph G_m via $\hat{F}^{hq}_{\{t\}}$, then pass G_m to the GCN to extract motion-related information. Finally, we extract audio feature via decoder to concatenate with the first layer of decoder, then fuse all the extracted features of same level and propagate to higher resolution layers.

During training stage, we first apply $L2$ loss between predicted keypoints K^{pred} and real keypoints K^{real} in Cartesian coordinate space:

$$\mathcal{L}^{Kpts}_{Coor} = \| K^{pred}_{\{1:P\}} - K^{real}_{\{1:P\}} \|_2, \tag{2}$$

We calculate the second loss in visual space with L1 loss:

$$\mathcal{L}^{Kpts}_{Vis} = \| H^{pred} - H^{real} \|_1 . \tag{3}$$

By applying the two losses in (Eq. (2) and Eq. (3)), we can obtain the loss of the predicted keypoints in both Cartesian coordinate space and spatial visual space which improve the keypoints prediction and facilitate the further video generation.

B. Coarse Video Generation: CVG

Despite of the local spatial information, the global appearance information, which can maintain the context of the video, is also crucial in generation. Therefore, we propose a coarse video generator (CVG) to simultaneously generate the general body appearance within each frame and smooth transition between adjacent frames from give music clip.

As shown in Fig. 2 (b), CVG consists of an AudioEncoder, an ImageEncoder, and an ImageDecoder. AudioEncoder processes audio sequence $A_{\{1:n\}}$ into audio features $F^{a}_{\{1:n\}}$ then feed to LSTM to obtain temporal information. ImageEncoder contains the top five layers of pretrained VGG network [35] and two additional convolution layers. In order to improve the continuity in the motion, we feed previous generated frame to the ImageEncoder to extract image feature $F^{v}_{\{t-1\}}$. Finally, we concatenate $F^{a}_{\{t\}}$ and $F^{v}_{\{t-1\}}$ along with a random variable z to ImageDecoder to obtain the current coarse video frame $V^{coarse}_{\{t\}}$.

Since we only expect the coarse video generation at this stage, we simply employ $L1$ loss between the real video frame $V^{real}_{\{t\}}$ and the generated coarse video frame $V^{coarse}_{\{t\}}$ for reconstruction:

$$\mathcal{L}^{vid}_{coarse} = \frac{1}{n}\sum_{t=1}^{n} \| V^{real}_{\{t\}} - V^{coarse}_{\{t\}} \|_1 . \tag{4}$$

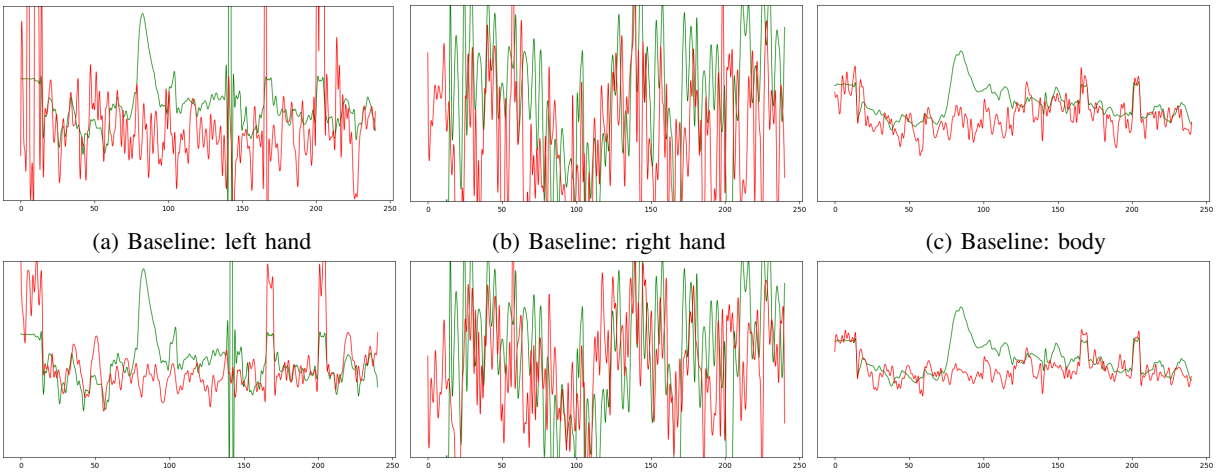
The output coarse videos can provide the general appearance information for the final high-quality video generation. Therefore, concatenate generated V^{coarse} and H^{pred} to feed to the next stage.

C. Final High-quality Video Generation: FHVG

To capture both intra-frame structure information and inter-frame temporal consistency, we propose a Structured Temporal UNet (STU) by leveraging the middle level information (the predicted keypoints K^{pred} and generated coarse video V^{coarse}) for final high-quality video generation, as shown in Fig. 3.

Firstly, we employ UNet [11] as our basic network, which is a prevalent network in image-to-image translation due to its ability of propagating context features from lower layers to higher resolution layers. However it ignores inter-frame temporal consistency, and suffers a jitter problem while synthesizing videos [10]. Herein, we propose to further temporally propagate a high-level feature between adjacent frames through the gated unit, then obtain the fused feature similar as GRU [36], but replacing the FC layers by CNNs to preserve spatial information. We refer to it as CNN-GRU in our paper.

Furthermore, conventional UNet only contains CNNs to extract features in spatial-level, while neglecting the intrinsic structure information. Therefore, we propose to explore the



(a) Baseline: left hand (b) Baseline: right hand (c) Baseline: body

(d) Ours: left hand (e) Ours: right hand (f) Ours: body

Fig. 4: Visualization of cello keypoints, where X-axis and Y-axis denote each sample and the 1-D PCA feature respectively. The red line and green line indicate the PCA features of predicted and ground truth keypoints respectively.

intra-frame structured between the motion components (the feature blocks located by the predicted keypoints) via GCN due to its ability of encoding the discrete features with the intrinsic structure.

The graph of motion components can be represented as $G_m = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V}, \mathcal{E}, \mathcal{A}$ denote the nodes, edges, and adjacency matrix of the graph respectively. The nodes of the graph are the feature blocks $\hat{F}^{hq}_{\{t\}}$ located by keypoints coordinates, and the edges are connected in the same manner as performed in OpenPose [37]. Then we feed G_m into GCN to aggregate this intra-frame local features to preserve the structure relationship during final generation.

Finally, we feed the keypoints, the heatmap concatenated coarse video, together with the given audio into the proposed STU to capture both intra-frame structure and temporal consistency for final video generation. An additional Audio-Video Discriminator is introduced to distinguish whether the given audio and video are paired. The STU and Audio-Video Discriminator therefore formed as a GAN [15], STU tries to fool the discriminator while the discriminator attempts to find the unpaired audio and video frames. The adversarial loss is:

$$\mathcal{L}^{G}_{hq} = \mathbb{E}[log(D(G(A, V^{coarse}, K^{pred}), A)], \tag{5}$$

and discriminator is trained with:

$$\begin{aligned}\mathcal{L}^{D}_{hq} =& \mathbb{E}[log(D(V^{real}, A)] + \\ & \mathbb{E}[log(D(G(A, V^{coarse}, K^{pred})), A)].\end{aligned} \tag{6}$$

Instead of simply using L1 loss in coarse video generation, we use perceptual loss [38] to capture high-level differences between the generated and real videos:

$$\mathcal{L}^{perc}_{hq} = \frac{1}{n}\sum_{i=1}^{n} \parallel \psi(V^{real}_{\{i\}}) - \psi(V^{hq}_{\{i\}}) \parallel_1, \tag{7}$$

where ψ denotes the output of different VGG-19 layers.

IV. Experiments

We evaluate our model on Sub-URMP [39] dataset to demonstrate the effectiveness of our proposed method for APVG task, followed by a detailed ablation study on each component and comparing our STU against other state-of-the-art video-to-video generation models.

A. Dataset and Implementation Details.

Sub-URMP [39] dataset consists of 13 instrument categories. Each category includes the performance videos of music clips recorded by 1 to 5 different people. In our experiments, we choose *cello* and *trombone* categories which contain 8000+ frames per person in the training set. We crop each frame into a square and resize to 256*256. Audios are extracted into Constant-Q transform (CQT) features [40] at the sampling rate of 44100Hz and hop length of 256 while each feature has a size of 84*87.

We adopt Adam optimizer with the learning rate starting from 0.001 and then gradually decreasing to 0.000125 during training. All the parameters in networks are initialized with Kaiming initialization [41].

B. Ablation Study

We first evaluate the contribution of each component in our method. We evaluate the qualitative result by the prevalent metrics: Peak Signal to Noise Ratio (PSNR), Structure Similarity Index Measure (SSIM) [42]. Table I reports the ablation study result, from which we can see, (1) All the three components, KHP, CVG and STU play important roles in our method. (2) By removing the GCN model (d) or CNN-GRU module (e) from STU, both PSNR and SSIM increase, which indicates their contributions.

C. Evaluation on Keypoints Predictor (KP)

To evaluate the performance of our keypoints prediction. We calculate the L2 distance between the predicted keypoints

Fig. 5: The generation examples of our model. Note that the results of vid2vid [10] was with the size of 144*256, which was resized into 256*453 for better formatting.

TABLE I: Quantitative evaluation of the proposed performance video generation. (Ours = Baseline + KHP+ CVG + STU, STU = GCN + CNN-GRU.)

Methods	Cello		Trombone	
	PSNR	SSIM	PSNR	SSIM
(a) w/o CVG	15.073	0.306	13.563	0.206
(b) w/o KHP	15.191	0.465	14.753	0.305
(c) w/o STU	15.767	0.536	14.656	0.362
(d) w/o GCN	16.253	0.551	15.572	0.395
(e) w/o CNN-GRU	16.437	0.548	15.519	0.395
Ours	**17.073**	**0.563**	**15.910**	**0.397**

and the real keypoints of the proposed KP together with its two variants As reported in Table II, the distance increased after removing the average condition ($avg.$ condition) or the differentiable landmark transformation (DLT), which verifies the contribution of each component. We further notice that the distance of trombone videos is much larger than that of cello, the reason is that OpenPose [37] fails more frequently to detect the ground truth keypoints on Trombone videos than on Cello ones, which affects the final video generation.

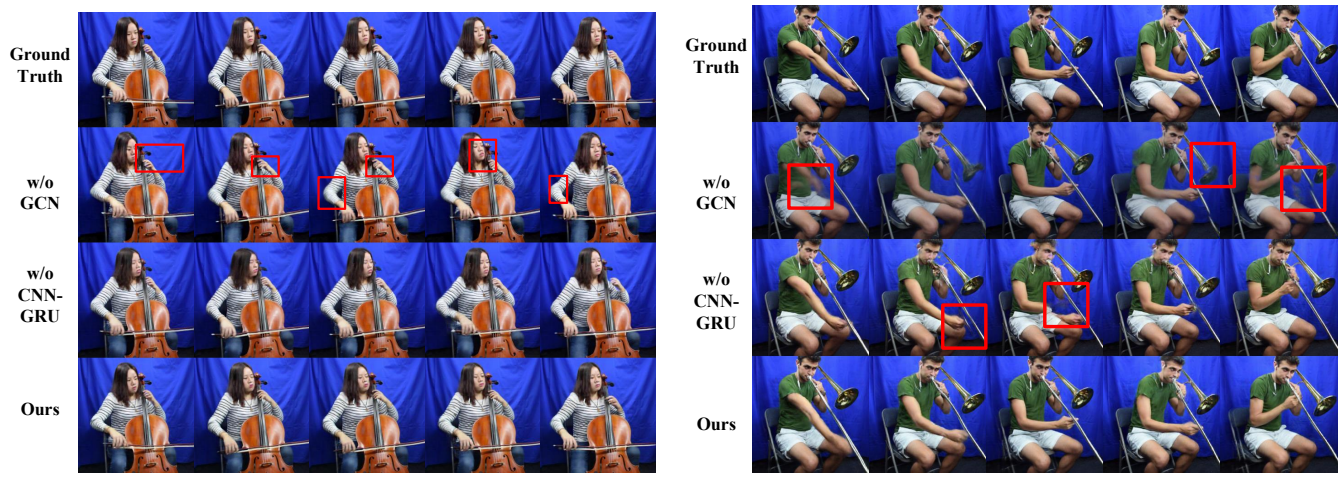
TABLE II: Evaluation different components in keypoints prediction. (Ours = Baseline + $avg.$ condition + DLT)

Methods	Mean Keypoints Distance		
	Cello	Trombone	Mean
Baseline	0.164	0.598	0.381
+ $avg.$ condition	0.151	0.427	0.289
+ DLT	0.152	0.490	0.321
Ours	**0.117**	**0.392**	**0.254**

Fig. 4 visualizes the turbulence between our predicted keypoints and the ground truth comparing to the baseline. It is clearly that our keypoint predictor can predict smoother and preciser keypoints than baseline (without average condition and the differentiable transformation). Note that the extremely large or small ground truths indicate the failure detection of OpenPose [37].

D. Evaluation on Structured Temporal UNet (STU)

As the first task of performance video generation, we leverage the predicted heatmap (transformed from landmark) as input and compare our proposed STU (video-to-video) with



(a) Comparison of cello results.

(b) Comparison of trombone results.

Fig. 6: The quality of generation with different experimental setting.

TABLE III: Quantitative evaluation of the proposed performance video generation with state-of-the-arts.

Methods	Cello		Trombone	
	PSNR	SSIM	PSNR	SSIM
EBDN [3]	13.553	0.246	12.358	0.225
vid2vid [43]	13.284	0.331	9.600	0.204
Ours	**17.073**	**0.563**	**15.910**	**0.397**

other state-of-the-art video generation methods, EveryBody Dance Now (EBDN) [3] and vid2vid [43]. As reported in Table III, our STU significantly beats the state-of-the-art video generation methods in all metrics. Fig. 5 further demonstrates two comparison examples on cello and trombone categories respectively. From Fig. 5, we can find that our predicted heatmap are more synchronized with the motions of ground truth and our synthesized coarse video contains the basic texture and the poses. The final generated high-quality video has the comparable quality to the ground truth.

E. User Study

We further provide a user study together with two examples to demonstrate the effectiveness of our model in Fig. 6 and Fig. 7. We first randomly select 12 video sets, each of which contains three videos generated by our method, our method without CNN-GRU and our method without GCN, then invite participants to vote on realistic and synchronism. Clearly, (1) our model achieves the highest rating than other variants in both realistic and synchronization. (2) Both the realistic and synchronization w/o CNN-GRU in Cello gain much lower rating than in Trombone. That means the CNN-GRU plays more important roles in Cello video generation by capturing the temporal consistency in slower motion videos (Cello). (3) GCN turns to play more important role in Trombone since the fast motion in videos (Trombone) affect less to keypoints prediction.

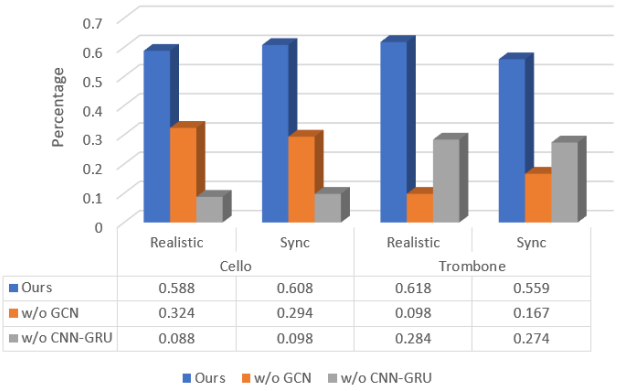


Fig. 7: User study of our model and its variants on both realistic and synchronization (Sync).

V. Conclusion

In this paper, we propose a novel multi-stage model for audio-driven performance video generation. To achieve this task, we first generate both global coarse video and local heatmap as middle information for final video generation. Then, we propose to transform keypoints to heatmap via a differentiable transforming function, since heatmap offers more spatial information while hard to generate from audio. Finally, a Structured Temporal UNet (STU) is designed to capture both intra-frame structured information via GCN module, and inter-frame temporal consistency via CNN-GRU based UNet module. Comprehensive experiments demonstrate the effectiveness of the proposed model.

Acknowledgment

This research is supported in part by Beijing Natural Science Foundation (Grant No. JQ18017), the National Natural Science Foundation of China (61976002), the Natural Science

Foundation of Anhui Higher Education Institutions of China (KJ2019A0033).

REFERENCES

[1] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *ACM Multimedia Conference on Multimedia*, 2018, pp. 627–635.

[2] L. Song, J. Cao, L. Song, Y. Hu, and R. He, "Geometry-aware face completion and editing," *arXiv preprint arXiv:1809.02967*, 2018.

[3] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," pp. 5933–5942, 2019.

[4] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Advances in Neural Information Processing Systems*, 2017, pp. 406–416.

[5] S. A. Jalalifar, H. Hasani, and H. Aghajan, "Speech-driven facial reenactment using conditional generative adversarial networks," *arXiv preprint arXiv:1803.07461*, 2018.

[6] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[7] J. Li, J. Yang, A. Hertzmann, J. Zhang, and T. Xu, "Layoutgan: Generating graphic layouts with wireframe discriminators," *arXiv preprint arXiv:1901.06767*, 2019.

[8] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio, "Obamanet: Photo-realistic lip-sync from text," *arXiv preprint arXiv:1801.01442*, 2017.

[9] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[10] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Advances in Neural Information Processing Systems*, 2018.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015, pp. 234–241.

[12] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8340–8348.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[14] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[16] S. Zhang, R. He, Z. Sun, and T. Tan, "Demeshnet: Blind face inpainting for deep meshface verification," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 637–647, 2017.

[17] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial discriminative heterogeneous face recognition," in *Conference on Artificial Intelligence*, 2018, pp. 7355–7362.

[18] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, "Deep audio-visual learning: A survey," *arXiv preprint arXiv:2001.04758*, 2020.

[19] L. Song, W. Wu, C. Qian, R. He, and C. C. Loy, "Everybody's talkin': Let me talk as you want," *CoRR*, vol. abs/2001.05201, 2020.

[20] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," *CoRR*, vol. abs/1807.07860, 2018.

[21] H. Zhu, A. Zheng, H. Huang, and R. He, "High-resolution talking face generation via mutual information approximation," *arXiv preprint arXiv:1812.06589*, 2018.

[22] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *European Conference on Computer Vision*, 2020.

[23] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," *arXiv preprint arXiv:1712.01393*, 2017.

[24] H. Zhou, Z. Liu, X. Xu, P. Luo, and X. Wang, "Vision-infused deep audio inpainting," in *International Conference on Computer Vision*, 2019, pp. 283–292.

[25] R. Fan, S. Xu, and W. Geng, "Example-based automatic music-driven conventional dance motion synthesis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 3, pp. 501–515, 2011.

[26] O. Alemi, J. Françoise, and P. Pasquier, "Groovenet: Real-time music-driven dance movement generation using artificial neural networks," *Networks*, vol. 8, no. 17, p. 26, 2017.

[27] J. Lee, S. Kim, and K. Lee, "Listen to dance: Music-driven choreography generation using autoregressive encoder-decoder network," *arXiv preprint arXiv:1811.00818*, 2018.

[28] E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman, "Audio to body dynamics," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[29] W. Zhuang, C. Wang, S. Xia, J. Chai, and Y. Wang, "Music2dance: Music-driven dance generation using wavenet," *arXiv preprint arXiv:2002.03761*, 2020.

[30] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[31] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz, "Dancing to music," in *Advances in Neural Information Processing Systems*, 2019.

[32] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics*, vol. 36, no. 4, p. 95, 2017.

[33] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable gans for pose-based human image generation," in *Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3408–3416.

[34] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "Unsupervised person image synthesis in arbitrary poses," in *Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8620–8628.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[36] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[37] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," in *arXiv preprint arXiv:1812.08008*, 2018.

[38] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016, pp. 694–711.

[39] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017, pp. 349–357.

[40] C. Schörkhuber and A. Klapuri, "Constant-q transform toolbox for music processing," in *Sound and Music Computing Conference*, 2010, pp. 3–64.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *International Conference on Computer Vision*, 2015, pp. 1026–1034.

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[43] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," *arXiv preprint arXiv:1808.06601*, 2018.