Cascade Attention Guided Residue Learning GAN for Cross-Modal Translation

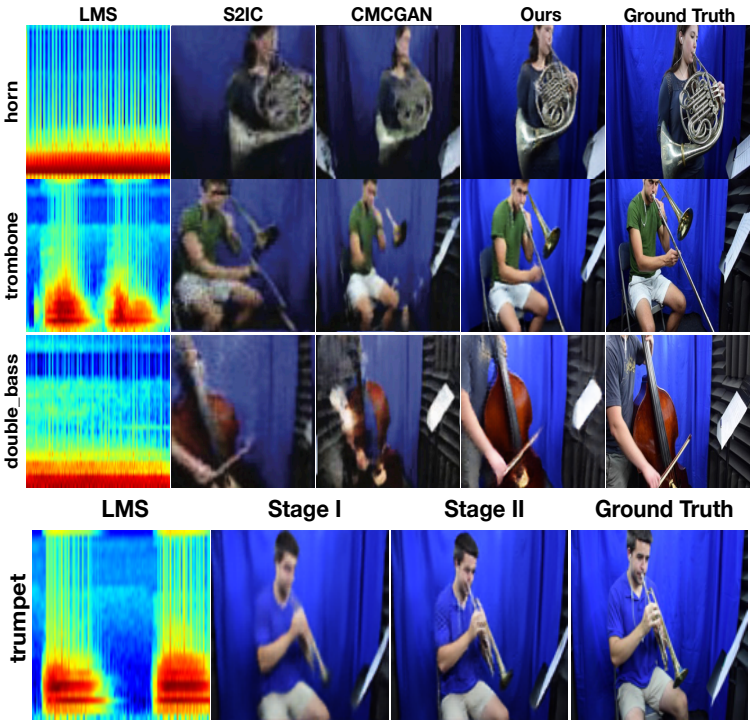
Bin Duan, Wei Wang, Hao Tang, Hugo Latapie, Yan Yan

Introduction

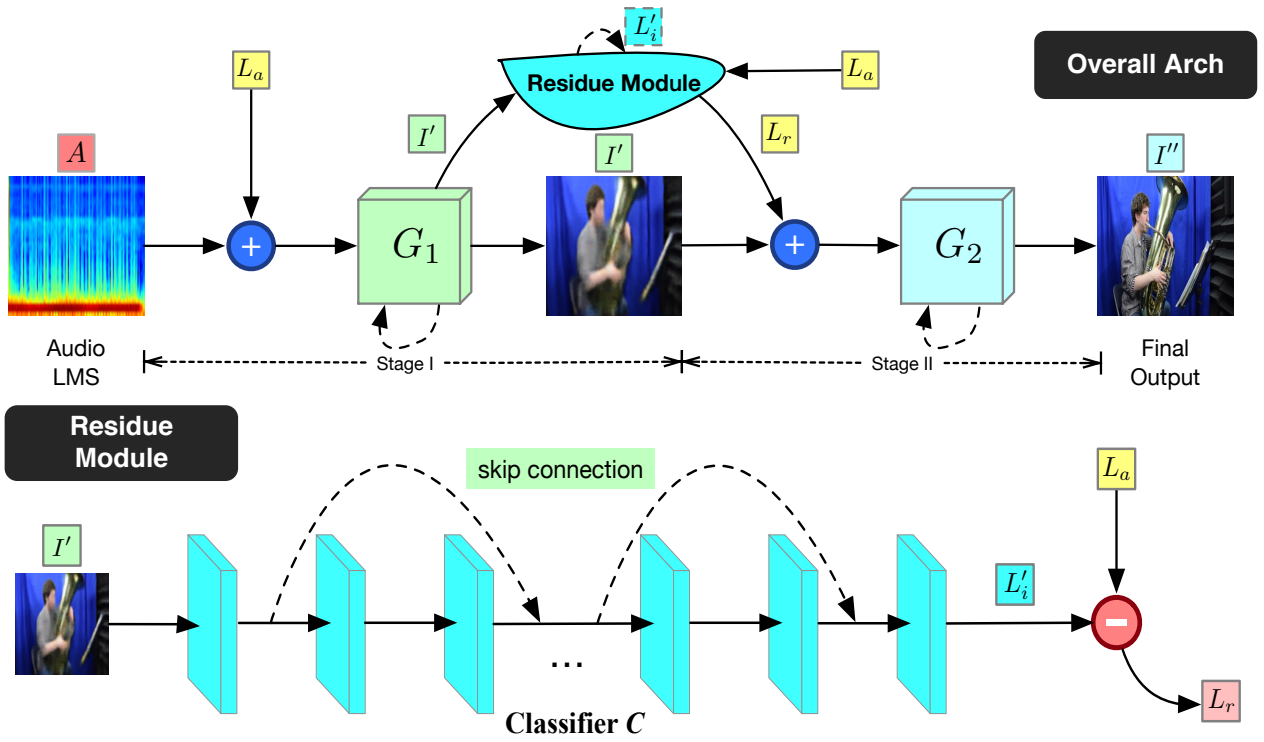
Problem 1: With a huge gap between audio and image modality, how to generate images from audios?

Problem 2: Single-stage generation network is not able to capture the complex relationships between the two modalities; thus it leads to suboptimal translation.

Qualitative Results



CAR-GAN Framework



Quantitative Results

Method	Accuracy	
	Training	Testing
S2IC	0.8737	0.7556
CMCGAN	0.9105	0.7661
Ours	**0.9954**	**0.9068**

CAR-GAN Formulation

$$\mathcal{L}(I', I'') = \lambda_{I'}\mathcal{L}(I') + \lambda_{I''}\mathcal{L}(I''),$$

$$\mathcal{L}_{cG_1}(A, I') = \mathbb{E}_{A,I}\left[\log D(A, I)\right] + \mathbb{E}_{A,I'}\left[\log(1 - D(A, I'))\right]$$

$$\mathcal{L}_{cG_2}(A, I'') = \mathbb{E}_{A,I}\left[\log D(A, I)\right] + \mathbb{E}_{A,I''}\left[\log(1 - D(A, I''))\right]$$

$$\min_{G_1,G_2,C}\max_{D}\mathcal{L} = \lambda_c\mathcal{L}_C + \lambda_{cG}\mathcal{L}_{cG} + \lambda_{L1}\mathcal{L}_{L1}$$

Reference

[1] L. Chen, S. Srivastava, Z. Duan, and C. Xu. Deep cross-modal audio-visual generation. In ACM MM Workshop, 2017.
[10] W. Hao, Z. Zhang, and H. Guan. CMCGAN: A uniform framework for cross-modal visual-audio mutual generation. In AAAI, 2018.
[46] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In ICCV, 2017.