DEEP VIDEO INPAINTING GUIDED BY AUDIO-VISUAL SELF-SUPERVISION

Kyuyeon Kim, Junsik Jung, Woo Jae Kim*, Sung-Eui Yoon*

School of Computing, Korea Advanced Institute of Science and Technology (KAIST)

ABSTRACT

Humans can easily imagine a scene from auditory information based on their prior knowledge of audio-visual events. In this paper, we mimic this innate human ability in deep learning models to improve the quality of video inpainting. To implement the prior knowledge, we first train the audio-visual network, which learns the correspondence between auditory and visual information. Then, the audio-visual network is employed as a guider that conveys the prior knowledge of audio-visual correspondence to the video inpainting network. This prior knowledge is transferred through our proposed two novel losses: audio-visual attention loss and audio-visual pseudo-class consistency loss. These two losses further improve the performance of the video inpainting by encouraging the inpainting result to have a high correspondence to its synchronized audio. Experimental results demonstrate that our proposed method can restore a wider domain of video scenes and is particularly effective when the sounding object in the scene is partially blinded.

***Index Terms*—** audio-visual learning, audio-visual correspondence, audio-visual network, deep video inpainting

1. INTRODUCTION

Imagine hearing the sound of a bird singing. You may come up with an image of a bird flying in the sky or sitting on top of a tree. In this fashion, humans can easily visualize a scene related to incoming auditory signals [1]. This natural behavior is empowered by the prior knowledge of semantic mapping between the visual and auditory modalities learned from ubiquitous audio-visual events around us. This ability to connect the dots between two modalities allows humans to restore videos better whose spatial information is corrupted. In other words, even though the video is partially blinded, humans can easily imagine what is happening in missing parts by listening to the corresponding audio. Based on this intuition, our work tries to mimic this human ability in deep learning models to better solve the following video inpainting problem: filling in missing visual regions in a video, guided by the audio signal. Hence, our goal can be articulated into answering the following question: can machines also learn to restore the visual content of a video by hearing its corresponding sound?

To achieve such goal, we exploit the audio-visual correspondence learned by the **audio-visual network (AV-Net)** [2] to train the **video inpainting network (VI-Net)**. The AV-Net learns to generate an audio-visual attention map that highlights visual regions which are corresponding to the synchronized audio and to capture the pseudo-class of each modality within the audio-visual pair. In this manner, the AV-Net learns the semantic relationship within the audio-visual pairs without video labels in a self-supervised manner, without labeled videos. There have been previous attempts to use this audio-visual correspondence for several of their unique downstream tasks, such as sounding object localization [2, 3] and sound source separation [4, 5]. Unlike these attempts, we aim to leverage the prior knowledge of audio-visual correspondence for the video inpainting task, which has not been explored yet.

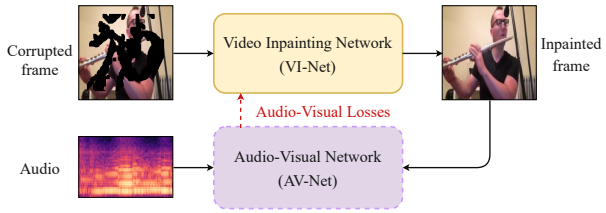

Fig. 1. Overview of our proposed method. We use the audio-visual network (AV-Net) as a guider of the video inpainting network (VI-Net) that conveys the prior knowledge of audio-visual correspondence through our proposed audio-visual losses.

As shown in Fig. 1, the AV-Net guides the VI-Net to use the corresponding audio signal as an important cue for restoring the corrupted frame. Given the prior information of audio-visual correlation that AV-Net provides, we propose two novel audio-visual losses to convey the prior knowledge to the VI-Net: **audio-visual attention loss** and **audio-visual pseudo-class consistency loss**. Audio-visual attention loss encourages the VI-Net to minimize the disparity of the audio-visual attention maps between the original and the inpainted frame. By doing so, the VI-Net solely focuses on restoring areas corresponding to the sounding object, making the inpainting result semantically more accurate. Audio-visual pseudo-class consistency loss is designed to indicate that visual and audio information from the same video should belong to the identical class. Using auxiliary classifiers, we encourage the VI-Net to learn that the visual features of inpainted frames and the synchronized audio features should belong to the same pseudo-classes. This audio-guided class consistency information can further enhance the video inpainting quality.

In summary, our main contributions are as follows:

- To enhance the video inpainting quality, we propose a novel approach that utilizes the inherent sound from the video itself.
- Based on the pretrained AV-Net, we propose two novel losses that enable the VI-Net to utilize the inherent sound of a video for restoring corrupted frames.
- Experimental results show that our approach is especially effective when restoring the frame whose sounding object in the scene is partially blinded.

2. RELATED WORK

In this section, we briefly discuss two research domains relevant to our work: deep video inpainting and audio-assisted visual synthesis.

* Equal contributions

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. 2019R1A2C3002833)

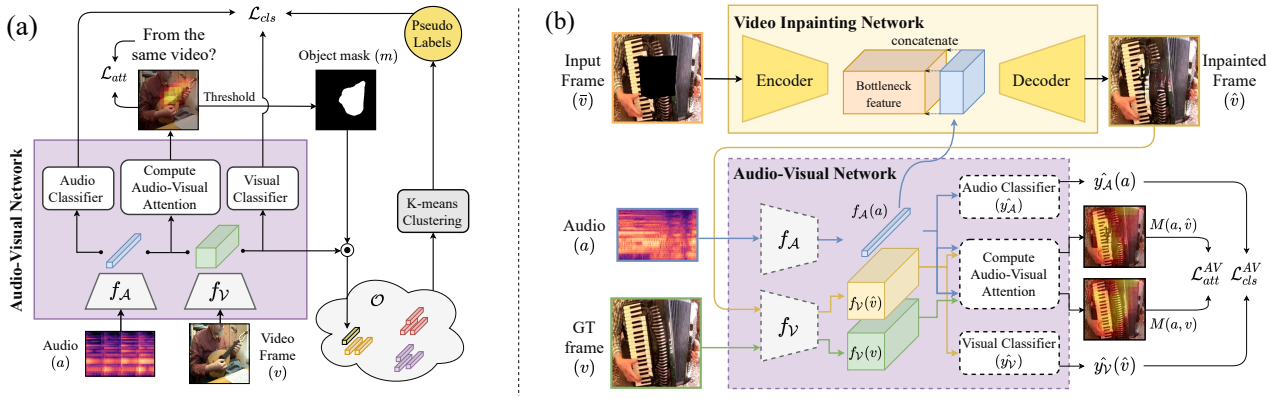

Fig. 2. Schematics of **(a) Training the audio-visual network (AV-Net)** and **(b) Training the video inpainting network (VI-Net)** with the audio-visual guidance from the pretrained AV-Net. Modules with dotted line indicate that their parameters are frozen in the training time.

Deep video inpainting. Video inpainting is a challenging problem aiming to restore missing regions in consecutive frames with spatially and temporally plausible content [6]. Recent approaches have achieved significant improvements via deep learning by using encoder-decoder-based architectures [7]. Along this line, there have been attempts of adopting optical flows [8, 9, 10], novel architectures [11, 12], attention modules [13, 14], and adversarial mechanisms [15]. Despite these successes, little attention has been given to employing the audio signal, which is the innate correspondence prior within a video. Thus, our work takes a pioneering step to demonstrate a generic method of utilizing audio signals to the video inpainting problem.

Audio-assisted visual synthesis. Audio has been used as an effective prior for synthesizing images or video frames, but in limited application domains. Such domains include audio-based adversarial image generation [16], talking face synthesis [17, 18], and speakers' face super-resolution [19]. Compared to these previous approaches, our work has the following distinctions. While [16] used human-labeled videos to obtain semantic knowledge, our work utilizes an audio-visual relationship learned from the self-supervised training procedure. We also consider a broader scope of audio-visual events occurring in the real world, rather than managing only the video of talking faces as in [17, 18, 19].

3. PROPOSED METHOD

In this section, we explain our audio-guided video inpainting framework. Fig. 2 shows an overview of this framework consisting of two main parts: the audio-visual network (AV-Net) and the video inpainting network (VI-Net). We first review the AV-Net (Sec. 3.1). Then, we provide details on two novel losses derived from the AV-Net that are used to train the VI-Net: audio-visual attention loss and audio-visual pseudo-class consistency loss (Sec. 3.2).

3.1. The audio-visual network

Let $\mathcal{X} = \{(a_i, v_j) \mid 1 \leq i \leq N, 1 \leq j \leq N\}$ denote a set of audio-visual pairs such that a pair (a_i, v_j) is sampled from N number of videos. Here, a and v each represents the audio signal and the video frame. Given the pair $(a_i, v_j) \in \mathcal{X}$ as input, we aim to obtain the prior information in two forms: audio-visual attention map and pseudo-class of each input's modality. The former considers a pair (a_i, v_j) where a_i and v_j are each randomly sampled from N videos, while the latter considers only a pair drawn from the same video (i.e., $i = j$). Our training methodology of the AV-Net refers to [2, 20].

Audio-visual attention map. As shown in Fig. 2 (a), the AV-Net consists of two convolutional sub-networks for feature extraction: audio network $f_\mathcal{A}$ and visual network $f_\mathcal{V}$. From two sub-networks, we extract an audio feature $f_\mathcal{A}(a_i) \in \mathbb{R}^c$ and a visual feature map $f_\mathcal{V}(v_j) \in \mathbb{R}^{h \times w \times c}$. Note that $h \times w$ and c denote spatial and channel dimensions, respectively. Then, we obtain the similarity map of $\mathbb{R}^{h \times w}$ by computing the scalar product between L_2-normalized $f_\mathcal{A}(a_i)$ and $f_\mathcal{V}(v_j)$ along the channel dimension for each of the spatial units within $f_\mathcal{V}(v_j)$. The similarity map then describes how strongly each spatial location of $f_\mathcal{V}(v_j)$ reacts to the audio descriptor $f_\mathcal{A}(a_i)$. Finally, we apply a sigmoid operation to this similarity map to obtain the audio-visual attention map $M(a_i, v_j) \in \mathbb{R}^{h \times w}$.

Intuitively, the attention map $M(a_i, v_j)$ would show high attention in the area that semantically corresponds to both the given audio a_i and the video frame v_j. Based on this intuition, the objective of training the AV-Net can be formulated into a binary classification problem as follows:

$$\mathcal{L}_{att} = \texttt{BCE}\left(y_{corr}, \texttt{GMP}(M(a_i, v_j))\right), \tag{1}$$

where $\texttt{BCE}(\cdot,\cdot)$ denotes the binary cross-entropy loss, $\texttt{GMP}(\cdot)$ denotes the global max-pooling operation, and y_{corr} denotes a binary label that indicates whether the audio-visual pair comes from the same video. By minimizing the cross-entropy between y_{corr} and the largest value of the attention map $M(a_i, v_j)$, the network is encouraged to maximize the attention values in regions that correspond to the given audio a_i, and to suppress them when audio-visual pairs do not match.

Pseudo-class prediction. Pseudo-labels of audio and visual features are also used to stabilize the training of the AV-Net. For pseudo-label extraction, we use matching audio-visual pairs $(a_i, v_i) \in \mathcal{X}$. We apply set threshold to the attention map $M(a_i, v_i)$ to obtain a binary mask $m_i \in \{0, 1\}^{h \times w}$. Using this attention-based binary mask, we compute the object representation $o_i \in \mathbb{R}^c$ from the visual feature $f_\mathcal{V}(v_i)$ to pick out the area where the audio-visual event is present. In specific, $o_i = \texttt{GAP}\left(m_i \odot f_\mathcal{V}(v_i)\right)$, where $\texttt{GAP}(\cdot)$ and $\odot$ denote the global average pooling operation and the channel-wise Hadamard product, respectively. We finally perform a K-means clustering on the set of object descriptors $\mathcal{O} = \{o_1, o_2, \cdots, o_N\}$ to assign each of them a pseudo-label corresponding to the cluster to which it belongs.

With these pseudo-labels set as ground truth, the network and

classifiers are trained by minimizing the following classification objective:

$$\mathcal{L}_{cls} = \texttt{CE}\left(y_p(o_i), \hat{y_{\mathcal{A}}}(a_i)\right) + \texttt{CE}\left(y_p(o_i), \hat{y_{\mathcal{V}}}(v_i)\right), \quad (2)$$

where $\texttt{CE}(\cdot,\cdot)$ denotes the categorical cross-entropy loss between two logit vectors and $y_p(o_i)$ represents the one-hot pseudo-label of o_i. $\hat{y_{\mathcal{A}}}(a_i)$ and $\hat{y_{\mathcal{V}}}(v_i)$ indicate the logit vectors from the linear classifiers $\hat{y_{\mathcal{A}}}$ and $\hat{y_{\mathcal{V}}}$ given a_i and v_i, respectively.

We train the AV-Net with Eq. 1 and Eq. 2 in an alternate manner, as two objectives mutually improves the overall performance [21].

3.2. Training the video inpainting network

In this sub-section, we describe our two novel losses derived from the AV-Net to further improve the training of the VI-Net. Suppose a pair of a corrupted video frame $\bar{v}$ and its ground truth frame v. Then, the VI-Net returns an inpainted frame $\hat{v}$ given the corrupted frame $\bar{v}$.

Audio-visual attention loss. We exploit the ability of the AV-Net to localize the sounding object in order to design our novel audio-visual attention loss. The audio-visual network takes video frame v and its paired audio a as inputs and generates an attention map $M(a, v)$ which highlights the area matching the given audio a. In the same way, the attention map $M(a, \hat{v})$ can be obtained by replacing v with $\hat{v}$. The key idea is that if the spatial content of the audio-visual event are successfully recovered in $\hat{v}$, the attention maps $M(a, \hat{v})$ and $M(a, v)$ should be identical. Otherwise, $M(a, \hat{v})$ would be vastly different from $M(a, v)$, especially in the area where the audio-visual event takes place.

From the investigation above, we observe that minimizing the difference between these two attention maps would reduce the disparity between v and $\hat{v}$. Hence, we propose the following audio-visual attention loss:

$$\mathcal{L}_{att}^{AV} = \frac{1}{hw} \left\| M(a, v) - M(a, \hat{v}) \right\|_2^2, \quad (3)$$

where h and w are the height and width of $M(\cdot,\cdot)$, respectively. This objective encourages the VI-Net to reconstruct the corrupted frame in a way such that the audio-visual attention map of the inpainted frame $M(a, \hat{v})$ is similar to the attention map of the ground truth frame $M(a, v)$. As a result, the inpainting network can better restore the missing part within sound-salient areas by filling it with content or texture that actively reacts to the given audio feature. This property cannot be found in common reconstruction losses (e.g., L_1 loss), which ignore additional cues from the audio.

Audio-visual pseudo-class consistency loss. To further improve the performance of the VI-Net, we additionally guide it with the class-consistency information between the audio and video frame inputs. The audio and visual information from a synchronized video should semantically belong to the same class. Hence, by learning that the restored frame $\hat{v}$ should belong to the same class as the corresponding audio a, the VI-Net could better reconstruct $\hat{v}$ such that it is more similar to the ground truth frame v.

We inject the audio information to the VI-Net by concatenating the audio feature $f_{\mathcal{A}}(a)$ to the bottleneck feature from the encoder of the VI-Net (the upper part of Fig. 2 (b)). Note that we broadcast the audio feature $f_{\mathcal{A}}(a)$ to the spatial dimension of the bottleneck feature before the concatenation. As the pretrained AV-Net can already predict the pseudo-class of the audio a, we set this as a guideline to determine whether the inpainted frame $\hat{v}$ has coherent content. Therefore, we design the audio-visual pseudo-class consistency loss as follows:

$$\mathcal{L}_{cls}^{AV} = \texttt{CE}\left(\hat{y_{\mathcal{A}}}(a), \hat{y_{\mathcal{V}}}(\hat{v})\right), \quad (4)$$

where $\hat{y_{\mathcal{A}}}(a)$ and $\hat{y_{\mathcal{V}}}(\hat{v})$ denote the logit vectors from the linear classifiers given a and $\hat{v}$, respectively. Note that the linear classifiers are also pretrained and frozen as parts of the AV-Net. Audio-visual pseudo-class consistency loss guides the VI-Net to synthesize a frame $\hat{v}$ that is class-consistent with the synchronized audio a.

Total loss. To train the VI-Net, we use the final loss as follows:

$$\mathcal{L} = \lambda_{L_1}\mathcal{L}_{L_1} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{att}^{AV}\mathcal{L}_{att}^{AV} + \lambda_{cls}^{AV}\mathcal{L}_{cls}^{AV}, \quad (5)$$

where $\mathcal{L}_{L_1}$ and $\mathcal{L}_{adv}$ respectively denote the L_1 loss and the adversarial loss from T-PatchGAN [15]. Note that these two losses are borrowed from [13], which is our baseline VI-Net. The VI-Net is optimized jointly with our proposed losses $\mathcal{L}_{att}^{AV}$ and $\mathcal{L}_{cls}^{AV}$. Hence, the network learns to consider the audio-visual consistency while minimizing the visual difference. The weights for each loss are empirically set as follows: $\lambda_{L_1} = 1$, $\lambda_{adv} = 0.01$, $\lambda_{att}^{AV} = 2$, and $\lambda_{cls}^{AV} = 1$.

4. EXPERIMENTS

4.1. Experimental settings

Datasets. We adopt **AVE** [22] and **MUSIC-Solo** [4] dataset to show the effectiveness of our approach. AVE dataset contains 4113 video clips covering 29 categories of diverse real-life audio-visual events. MUSIC-Solo dataset contains 493 video clips with 11 categories that exclusively cover solo performances of diverse musical instruments. We follow the official split of AVE dataset. On the other hand, we randomly split MUSIC-Solo dataset into 343/50/100 for train/validation/test since there is no designated split.

Moreover, we evaluate our method on two types of maskings: **I-mask** and **S-mask**. I-masks irregularly blind the pixels with random strokes and shapes. We adopt the subset of NVIDIA Irregular Mask Dataset [23]. For testing, we randomly pick three masks with a blinding ratio of 20.0%, 27.7%, and 28.4%, respectively. We also design S-masks to blind the region which corresponds to the sounding object. We collect S-masks by eroding the object mask m_i mentioned in Sec. 3.1 until the spatial area of the masking covers 20% of the image. This ratio refers to the approximate proportion of the region that the sounding object occupies in the video scene.

Preprocessing. Given a video clip, we extract video frames at 8 fps and resample its mono-channel audio at 16 kHz. Then, the video frame is resized to the spatial size of 256×256 and then randomly cropped (for training) or resized (for testing) into 224×224. The audio is sampled by retrieving a 1-second segment, and converted to the log-scale mel spectrogram with 0.01-second window size, half-window hop length, and 80 mel bins, finally treated as a single-channel matrix with the spatial dimension of 201×80.

Audio-visual network We follow [2, 20] to implement the audio-visual network (AV-Net). For visual and audio sub-networks $f_{\mathcal{V}}$ and $f_{\mathcal{A}}$, we use ResNet-18-based architectures as in [20].

Video inpainting baseline. We adopt one of the state-of-the-art architectures, the Spatial-Temporal Transformer Network (STTN) [13] as our baseline model. As our major interest lies in inpainting videos with audio-visual events, our choice of video dataset is different from the original work [13]. Therefore, we train the STTN on the aforementioned datasets from scratch, without audio signals.

Training details. To train the AV-Net, we adopt Adam optimizer with the learning rate of `5e-5` for AVE and `1e-4` for MUSIC-Solo dataset. The batch size is set to 32 for both datasets. Furthermore, we set the threshold value to 0.07 while obtaining binary masks and the number of clusters to 10 while collecting the pseudo-classes of

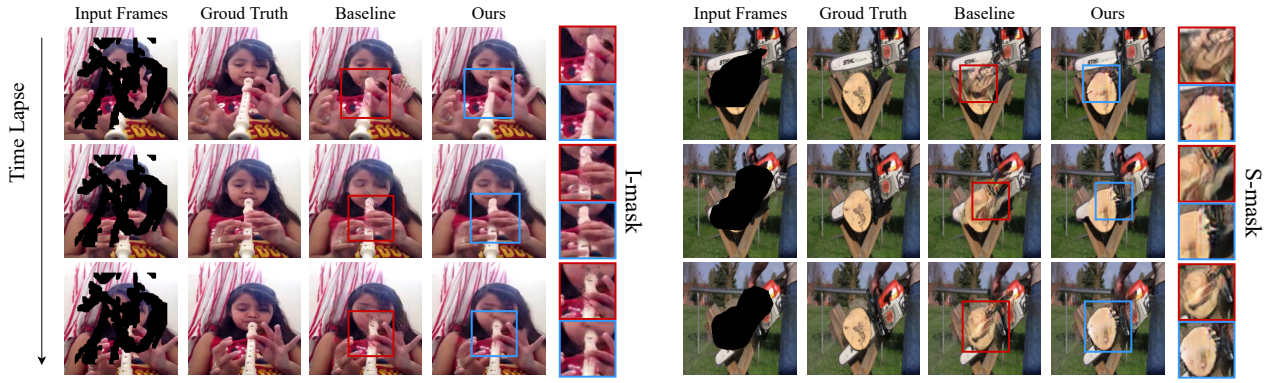

Fig. 3. Qualitative results of two samples from AVE dataset blinded by **I-masks** (left) and **S-masks** (right). While the baseline STTN without audio signals shows more artifacts around the object and produces blurry result, our method produces more realistic and clearer results.

object representations. While training the AV-Net for 4 epochs total, the learning rate is decayed by 0.1 after 2 epochs. Then, the parameters of the pretrained AV-Net are frozen while training the VI-Net. For AVE dataset, we train the VI-Net using Adam optimizer with the initial learning rate of `1e-4` decayed by 0.1 for every 100k iterations for a total of 350k iterations. For MUSIC-Solo dataset, due to the lack of training data, we fine-tune the VI-Net pretrained on AVE dataset using Adam optimizer for a total of 100k iterations with the learning rate of `1e-5` for first 50k iterations, and `1e-6` for the remaining iterations. For both datasets, the batch size is set to 8.

Evaluation metrics. The quantitative result is reported using three widely-used metrics: PSNR [9], SSIM [24], and video-based Fréchet Inception Distance (VFID) [15]. In detail, PSNR and SSIM are standard metrics to assess the synthesized scenes, whereas VFID quantifies the perceptual difference compared to the ground truth.

Method			I-mask			S-mask		
	$\mathcal{L}_{att}^{AV}$	$\mathcal{L}_{cls}^{AV}$	PSNR↑	SSIM↑	VFID↓	PSNR↑	SSIM↑	VFID↓
Baseline	-	-	30.76	93.45	3.549	26.58	91.93	5.553
w/ Ours	-	✓	30.81	93.55	3.356	26.83	92.21	5.305
w/ Ours	✓	-	30.94	93.61	3.273	27.16	92.47	5.271
w/ Ours	✓	✓	**31.18**	**93.65**	**3.184**	**27.32**	**92.69**	**4.961**

Table 1. Quantative evaluation and ablation study of applying our method on **AVE** dataset with two different types of masks. ↑ indicates that higher is better and ↓ means that lower is better.

Method			I-mask			S-mask		
	$\mathcal{L}_{att}^{AV}$	$\mathcal{L}_{cls}^{AV}$	PSNR↑	SSIM↑	VFID↓	PSNR↑	SSIM↑	VFID↓
Baseline	-	-	29.49	93.85	4.316	26.47	91.88	5.706
w/ Ours	-	✓	29.60	93.84	4.191	26.95	92.38	5.205
w/ Ours	✓	-	29.66	**93.87**	4.221	27.05	92.40	5.194
w/ Ours	✓	✓	**29.68**	93.84	**4.092**	**27.12**	**92.53**	**4.929**

Table 2. Quantative evaluation and ablation study of applying our method on **MUSIC-Solo** dataset with two different types of masks. ↑ indicates that higher is better and ↓ means that lower is better.

4.2. Result and discussion

We test our method on 4 different experimental setups derived from the combinations of video and mask datasets mentioned in Sec. 4.1. Table 1 shows that adopting our proposed audio-visual objectives outperforms the visual-only baseline on AVE dataset for all suggested metrics. As shown in Table 2, our method also performs substantially well on MUSIC-Solo dataset with video scenes strictly related to musical instruments. Performance improvements over two different video datasets also show that our method is effective not only in domain-specific videos such as MUSIC-Solo dataset but also in videos with a broader domain such as AVE dataset. Ablation studies in Table 1 and 2 imply that our two losses harmoniously give a positive impact on the inpainting quality, with the audio-visual attention loss showing a bigger influence.

One interesting point is that performance gains on S-masks are larger than those on I-masks. As shown in Table 1, on AVE dataset masked with I-masks, our method of applying both audio-visual losses improves the baseline PSNR and VFID by 0.42 and 0.365, respectively. On the same dataset with S-masks, our method shows a larger PSNR increase of 0.74 and VFID improvement of 0.592. The same tendency is shown in Table 2 on the MUSIC-Solo dataset. In the case of I-masks, PSNR and VFID improvement each shows 0.19 and 0.224 compared to the baseline. On the other hand, improvements are greater in the case of S-masks, showing PSNR increase of 0.65 and VFID improvement of 0.777. Recalling that S-masks are designed to mask audio-visual events, this tendency indicates that our method indeed effectively restores those regions. This shows that the audio-visual correspondence given as the prior information allows the video inpainting model to better restore regions corresponding to audio-visual events.

Fig. 3 demonstrates that our method produces more pleasing results for both types of masking. While the baseline model produces blurry artifacts around the sounding object, our approach can synthesize plausible results. Particularly, when the audio-visual event is partially deteriorated (by S-masks), the baseline fails to generate a realistic scene in the blinded area. In contrast, our method successfully restores the frame with clearer and comprehensible content while preserving the audio-visual coherency.

5. CONCLUSION

In this paper, we investigate a novel approach to using audio for video inpainting tasks by employing audio-visual self-supervision. We adopt the audio-visual network to bridge the gap between visual and audio modality, then use its functionalities to further guide the video inpainting network through proposed two novel losses: audio-visual attention loss and audio-visual pseudo-class consistency loss. Experimental results on two different audio-visual datasets – AVE and MUSIC-Solo dataset – with two types of masking – I-mask and S-mask – show that our approach improves the performance of the video inpainting network compared to the baseline.

6. REFERENCES

[1] Relja Arandjelovic and Andrew Zisserman, "Look, listen and learn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 609–617.

[2] Relja Arandjelovic and Andrew Zisserman, "Objects that sound," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 435–451.

[3] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman, "Localizing visual sounds the hard way," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16867–16876.

[4] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba, "The sound of pixels," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 570–586.

[5] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, 2018.

[6] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon, "Deep video inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5792–5801.

[7] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang, "Video inpainting by jointly learning temporal structure and spatial details," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019, vol. 33, pp. 5232–5239.

[8] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf, "Temporally coherent completion of dynamic video," *ACM Transactions on Graphics*, vol. 35, no. 6, pp. 1–11, 2016.

[9] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy, "Deep flow-guided video inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3723–3732.

[10] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf, "Flow-edge guided video completion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 713–729.

[11] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu, "Learnable gated temporal shift module for deep video inpainting," in *British Machine Vision Conference (BMVC)*, 2019.

[12] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim, "Onion-peel networks for deep video completion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4403–4412.

[13] Yanhong Zeng, Jianlong Fu, and Hongyang Chao, "Learning joint spatial-temporal transformations for video inpainting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 528–543.

[14] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim, "Copy-and-paste networks for deep video inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4413–4421.

[15] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu, "Free-form video inpainting with 3d gated convolution and temporal patchgan," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9066–9075.

[16] Chia-Hung Wan, Shun-Po Chuang, and Hung-Yi Lee, "Towards audio to scene image synthesis using generative adversarial network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 496–500.

[17] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman, "You said that?: Synthesising talking faces from audio," *International Journal of Computer Vision (IJCV)*, vol. 127, no. 11, pp. 1767–1779, 2019.

[18] Alexandros Koumparoulis, Gerasimos Potamianos, Samuel Thomas, and Edmilson da Silva Morais, "Audio-assisted image inpainting for talking faces," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7664–7668.

[19] Givi Meishvili, Simon Jenni, and Paolo Favaro, "Learning to have an ear for face super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1364–1374.

[20] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou, "Discriminative sounding objects localization via self-supervised audiovisual matching," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

[22] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 247–263.

[23] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[24] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.