석사학위논문
Master's Thesis

시청각적 자기지도를 통한 심층 비디오 인페인팅

Deep Video Inpainting Guided by Audio-Visual Self-Supervision

2022

김규연 (金奎延 Kim, Kyuyeon)

한국과학기술원

Korea Advanced Institute of Science and Technology

석사학위논문

시청각적 자기지도를 통한 심층 비디오 인페인팅

2022

김규연

한국과학기술원

전산학부

시청각적 자기지도를 통한 심층 비디오 인페인팅

김 규 연

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2021년 12월 8일

심사위원장　윤 성 의　(인)
심 사 위 원　김 민 혁　(인)
심 사 위 원　홍 승 훈　(인)

Deep Video Inpainting Guided by Audio-Visual Self-Supervision

Kyuyeon Kim

Advisor: Sung-Eui Yoon

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Computer Science

Daejeon, Korea
December 8, 2021

Approved by

Sung-Eui Yoon
Professor of School of Computing

The study was conducted in accordance with Code of Research Ethics[1].

---

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MCS

김규연. 시청각적 자기지도를 통한 심층 비디오 인페인팅. 전산학부 . 2022년. 21+iv 쪽. 지도교수: 윤성의. (영문 논문)
Kyuyeon Kim. Deep Video Inpainting Guided by Audio-Visual Self-Supervision. School of Computing . 2022. 21+iv pages. Advisor: Sung-Eui Yoon. (Text in English)

초 록

사람은 경험적으로 얻은 시청각적 사건에 관한 사전 지식에 의거하여 청각적 정보와 관련된 장면을 떠올릴 수 있다. 본 논문에서는 이와 같은 사람의 능력을 딥 러닝 모델에 모방하여 비디오 인페인팅 품질을 향상시키는 방법에 대해 탐구한다. 전술한 시청각적 사전 지식을 구현하기 위해, 시각 및 청각 정보 간의 연관성을 학습하는 시청각 네트워크를 학습시킨다. 이 시청각 네트워크를 안내자로 활용하여, 비디오 인페인팅 네트워크에게 시청각적 일관성에 대한 사전 지식을 전달하게 된다. 앞서 언급한 사전 지식은 본 연구에서 새로이 제시하는 두 가지 손실 함수인 시청각 어텐션 손실 함수 및 시청각 의사-클래스 일관성 보존 손실 함수를 통해 전달된다. 두 손실 함수를 통해 비디오 인페인팅 네트워크는 손상된 프레임이 주어진 소리와 높은 일관성을 보이도록 복원시킨다. 본 연구에서 제시한 방법은 다양한 종류의 시청각 이벤트에 대한 비디오 프레임을 잘 복원하는 것은 물론, 소리를 발생시키는 물체가 부분적으로 가려진 비디오 프레임을 복원하는 경우 더욱 효과적이다.

핵 심 낱 말 딥 러닝, 시청각 학습, 시청각적 연관성, 시청각 네트워크, 심층 비디오 인페인팅

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 to learn 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 – that further improve the performance of the video inpainting network. These two losses encourage 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. This thesis is based on the author's original paper [1].

Keywords Deep learning, Audio-visual learning, Audio-visual correspondence, Audio-visual network, Deep video inpainting

Contents

Contents . . . ii
List of Tables . . . iii
List of Figures . . . iv

Chapter 1. Introduction 1

Chapter 2. Related Work 3
- 2.1 Self-Supervised Audio-Visual Learning . . . 3
- 2.2 Deep Video Inpainting . . . 3
- 2.3 Audio-Assisted Visual Synthesis . . . 3

Chapter 3. Training the Audio-Visual Network 5
- 3.1 Audio-Visual Attention Map . . . 5
- 3.2 Pseudo-Class Prediction . . . 6

Chapter 4. Training the Video Inpainting Network 8
- 4.1 Audio-Visual Attention Loss . . . 9
- 4.2 Audio-visual pseudo-class consistency loss . . . 9
- 4.3 Total loss . . . 10

Chapter 5. Experimental Setup 11
- 5.1 Dataset and Preprocessing . . . 11
- 5.2 Implementation Details . . . 12

Chapter 6. Result 13
- 6.1 Evaluation Metrics . . . 13
- 6.2 Analysis and Discussion . . . 13

Chapter 7. Conclusion 16

Bibliography 17

Acknowledgments in Korean 20

Curriculum Vitae in Korean 21

List of Tables

6.1 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. 13

6.2 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. 13

List of Figures

1.1 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 relationship. The prior knowledge is conveyed through our proposed two audio-visual losses to further improve the performance of the video inpainting network. 1

3.1 The schematic of training the **audio-visual network** (AV-Net). The AV-Net is alternately trained with **attention-learning task** $\mathcal{L}_{att}$ (Sec. 3.1) and **pseudo-label classification task** $\mathcal{L}_{cls}$ (Sec. 3.2). Attention-learning task can be articulated into a binary classification task, which is predicting whether the audio-visual pair comes from the same video. Pseudo-labels of audio-visual pairs are determined by the clustering result over the set of object representations $\mathcal{O} = \{o_1, o_2, \cdots, o_N\}$, derived from refining visual features with object masks. These pseudo-labels are utilized for optimizing the AV-Net with the classification task. 6

4.1 The overall schematic of training the **video inpainting network** (VI-Net) guided by the **audio-visual network** (AV-Net). The prior knowledge which the AV-Net has learned is transferred to the VI-Net through our **audio-visual attention loss** $\mathcal{L}_{att}^{AV}$ (Sec. 4.1) and **audio-visual pseudo-class consistency loss** $\mathcal{L}_{cls}^{AV}$ (Sec. 4.2). Both losses encourage the VI-Net to recover the frame whose content has high correspondence to the synchronized audio. **Audio-visual attention loss** is defined as a difference between the attention map of the inpainted frame $M(a, \hat{v})$ and that of the ground truth frame $M(a, v)$. **Audio-visual pseudo-class consistency loss** supports the visual contents to be class-consistent $\hat{y_{\mathcal{V}}}(\hat{v})$ to the pseudo-class of the concantenated audio feature $\hat{y_{\mathcal{A}}}(a)$. Components with dotted line indicates that their weights are frozen during training time. 8

6.1 Qualitative results of two samples from the AVE dataset blinded by **I-masks** (top) and **S-masks** (bottom). While the baseline STTN shows more artifacts around the sounding object and produces blurry contents, our method produces more realistic and clearer results with less artifacts. 15

Chapter 1. Introduction

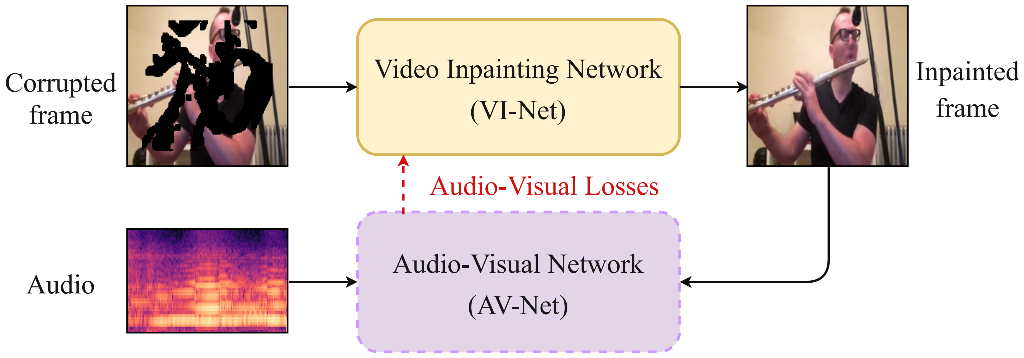
Corrupted frame
Video Inpainting Network (VI-Net)
Inpainted frame
Audio-Visual Losses
Audio
Audio-Visual Network (AV-Net)

Figure 1.1: 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 relationship. The prior knowledge is conveyed through our proposed two audio-visual losses to further improve the performance of the video inpainting network.

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 [2]. 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 a deep learning model to better solve the following video inpainting problem: a task of filling 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 scene in a video by hearing its corresponding sound?

The natural decision for this question might be combining the encoded audio with visual feature maps to make the audio act as conditional information. However, this approach has not always been successful, because the detailed semantic information within the audio often fails to be fused into vision domain directly [3]. Specifically, this strategy is available if the diversity of expected audio is limited, such as speech corpus [4, 5]. Nonetheless, it often fails to elicit low-level semantic information, such as gender and age within the speech-voice, making the audio signals negligible [6].

To achieve the goal while circumventing such issue, we exploit the audio-visual correspondence learned by the **audio-visual network** (AV-Net) [3] to train the **video inpainting network** (VI-Net). The AV-Net learns to generate an audio-visual attention map that highlights visual regions 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 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 [3, 7] and sound source separation [8, 9]. 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.

As shown in Fig. 1.1, 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 distill the prior knowledge to the VI-Net: **audio-visual attention loss** and **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 inpainted 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 performance.

In summary, this study proposes a method that utilizes the audio signal for the video inpainting task. We initially train the AV-Net in a self-supervised manner to implement the prior knowledge of the audio-visual relationship. This knowledge from the AV-Net is transferred to the VI-Net with two proposed losses, making the inpainted frame more realistic and semantically accurate. In other words, our contributions are as follows:

- For enhancing the video inpainting quality, we propose a novel approach that utilizes the inherent sound from the video itself (Ch. 4).
- Based on the pretrained AV-Net, we propose two novel losses – audio-visual attention loss and pseudo-class consistency loss – that enable the VI-Net to utilize the inherent sound from a video for restoring corrupted frames (Ch. 3, Ch. 4).
- Experimental results show that our approach is especially effective when restoring the frame whose sounding object in the scene is partially blinded (Ch. 6).

Chapter 2. Related Work

In this chapter, we briefly discuss three research domains relevant to our work: self-supervised audio-visual learning, deep video inpainting, and audio-assisted visual synthesis.

2.1 Self-Supervised Audio-Visual Learning

Even when videos are unlabeled, the innate correspondence between audio and visual stream provides useful supervision for audio-visual learning [2, 10]. This audio-visual correspondence has shown to be useful in designing the contrastive learning objectives [7, 11]. Learned audio-visual features have also been applied on several downstream tasks, such as sounding object localization [3, 7], speech or sound source separation [8, 9], and cross-modal synthesis guided by the modality of one another [12, 13, 14].

On the other hand, our work exploits audio-visual self-supervision for one of the visual restoration tasks – video inpainting – in order to restore highly semantic information in missing regions of video frames. We utilize this self-supervision to let the AV-Net learn the deep relationship between audio and visual feature, which will be further conveyed to support the better training of the VI-Net.

2.2 Deep Video Inpainting

Video inpainting is a challenging problem aiming to restore missing regions in consecutive frames with spatially and temporally plausible contents [15]. Recent methods have achieved significant improvements via deep learning by using encoder-decoder based architectures [16]. Among these learning-based methods, flow-based approaches utilize optical flow as a prior constraint for pixel-level propagation [17] or jointly train flow prediction and frame completion network to alleviate the blurriness [18, 19]. Instead of relying on additional prior, several studies introduce novel architectures [20, 21], attention modules [22, 23], or adversarial mechanisms [24] to ensure the spatio-temporal coherence.

Despite the success brought by these methods, little attention has been given to employing the audio signal, which is the innate correspondence prior of a video. Our work takes a pioneering step to demonstrate a generic method of utilizing audio signals to the video inpainting problem.

2.3 Audio-Assisted Visual Synthesis

Audio has been used as an effective prior for synthesizing image or video, but in limited application domains. Wan *et al.* [25] propose a GAN-based image generation framework that is conditioned by an audio feature vector but requires a human-labeled dataset. Jamaludin *et al.* [4] and Koumparoulis *et al.* [5] utilize audio features to synthesize a talking face of the given speech-voice and the speaker's identity. Meishvili *et al.* [6] design face super-resolution framework guided by gender and age information implicitly conveyed from the speech-voice.

Compared to these previous approaches, our work has the following distinctions. While [25] 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 than managing only the video of talking faces as in [4, 5, 6].

Chapter 3. Training the Audio-Visual Network

In this chapter, we review how the audio-visual network (AV-Net) embeds the prior knowledge of audio-visual events given diverse audio-visual pairs as input. The AV-Net first learns to quantify the coherency between the auditory information and the visual region within the video frame into a matrix called audio-visual attention map. This goal is achieved by training the AV-Net with an audio-visual correspondence task [3, 26], which is predicting whether the audio-visual pair belongs to the same video (Sec. 3.1). Leveraging the feature alignment effect from the previous task, the AV-Net accordingly learns to predict the pseudo-class of extracted audio and visual features by training the auxiliary classifier of each modality (Sec. 3.2). The overall structure of training the AV-Net is illustrated in Fig. 3.1.

Throughout this chapter, 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 the 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: the audio-visual attention map and the pseudo-class of input from each 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., $i = j$). Note that our training methodology of AV-Net refers to [3] and [27].

3.1 Audio-Visual Attention Map

As shown in Fig. 3.1, the AV-Net includes two convolutional sub-networks for feature extraction. Let (a_i, v_j) be an audio-visual pair that is arbitrary sampled from the aforementioned dataset $\mathcal{X}$. The audio network $f_{\mathcal{A}}$ takes a log-scale mel spectrogram a_i as an input and produces audio feature vector denoted as $f_{\mathcal{A}}(a_i) \in \mathbb{R}^c$. Furthermore, the visual network $f_{\mathcal{V}}$ extracts the feature map of $f_{\mathcal{V}}(v_j) \in \mathbb{R}^{h \times w \times c}$ from the input video frame v_j. Note that $h \times w$ and c denote spatial and channel dimensions, respectively. Based on these two feature maps, we obtain the similarity map of $\mathbb{R}^{h \times w}$ by computing the scalar product between $f_{\mathcal{A}}(a_i)$ and $f_{\mathcal{V}}(v_j)$ along the channel dimension for each of the spatial units within $f_{\mathcal{V}}(v_j)$. Note that both $f_{\mathcal{A}}(a)$ and $f_{\mathcal{V}}(v)$ are L_2 normalized to measure the cosine similarity from the scalar product. Then, the similarity map would describe 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 audio-visual attention map $M(a_i, v_j)$ would show high attention value in the area where the given audio a_i and the video frame v_j semantically correspond. Based on this intuition, the objective of training the AV-Net can be formulated into solving a binary classification problem, minimizing the loss function as follows:

$$\mathcal{L}_{att} = \texttt{BinaryCrossEntropy}\left(y_{corr}, \texttt{GlobalMaxPooling}(M(a_i, v_j))\right), \tag{3.1}$$

where y_{corr} denotes a binary label that indicates whether the components of the audio-visual pair comes from the same video or not (i.e., $y_{corr} = 1$ if $i = j$, and $y_{corr} = 0$ otherwise). By minimizing the

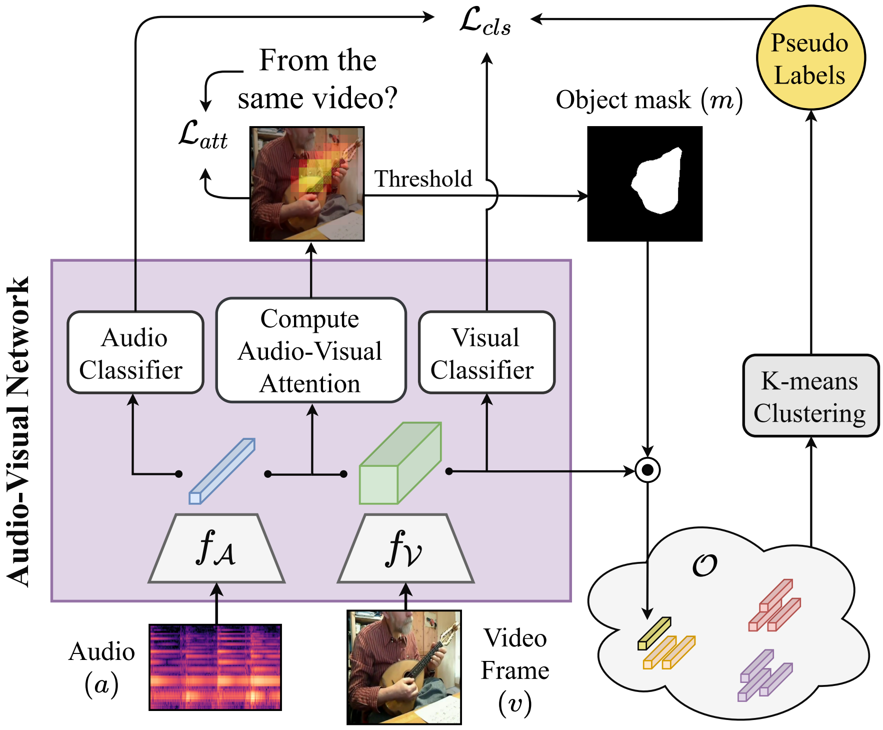

Figure 3.1: The schematic of training the **audio-visual network** (AV-Net). The AV-Net is alternately trained with **attention-learning task** $\mathcal{L}_{att}$ (Sec. 3.1) and **pseudo-label classification task** $\mathcal{L}_{cls}$ (Sec. 3.2). Attention-learning task can be articulated into a binary classification task, which is predicting whether the audio-visual pair comes from the same video. Pseudo-labels of audio-visual pairs are determined by the clustering result over the set of object representations $\mathcal{O} = \{o_1, o_2, \cdots, o_N\}$, derived from refining visual features with object masks. These pseudo-labels are utilized for optimizing the AV-Net with the classification task.

binary cross-entropy between y_{corr} and the largest value of the attention map $M(a_i, v_j)$ over the spatial dimension $h \times w$, the network is encouraged to maximize the attention values in regions that match to the given audio a_i while suppressing them when audio-visual pairs do not match. Further, we note that positive audio-visual pairs (where $i = j$) and mismatching pairs (where $i \neq j$) are sampled in equal ratio when training the AV-Net with the objective of Eq. 3.1.

3.2 Pseudo-Class Prediction

After training the AV-Net with the attention learning objective (Eq. 3.1), both audio and visual features are substantially aligned in the embedding space. This allows us to train a pseudo-class predictor (i.e., pseudo-label classifier) for each modality. Furthermore, pseudo-labels of both audio and visual features are additionally adopted to stabilize the training of the AV-Net. The detailed procedure of extracting pseudo-labels will be described below.

For pseudo-class extraction, we note that only the positive audio-visual pairs are used. In other words, we only consider the pairs satisfying $i = j$, which are $(a_i, v_i) \in \mathcal{X}$. With the audio-visual attention map $M(a_i, v_i)$ from the input pair, we apply set threshold to obtain a binary mask $m_i \in \{0, 1\}^{h \times w}$.

Accordingly, the true zone of m_i reveals the sounding object, while the false zone blinds the irrelevant backgrounds. 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 is gained by $o_i = \texttt{GlobalAveragePooling}\left(m_i \odot f_{\mathcal{V}}(v_i)\right)$, where the operator $\odot$ denotes the channel-wise Hadamard product. By masking out noisy and redundant information using the mask m_i, o_i contains more compact and distinctive information than $f_{\mathcal{V}}(v_i)$. 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. We emphasize that this clustering process is able to produce meaningful results by leveraging the feature alignment effect from the training objective $\mathcal{L}_{att}$ and noise-free o_i.

We append learnable linear classifiers to the end of each sub-network $f_{\mathcal{A}}$ and $f_{\mathcal{V}}$, which predicts pseudo-classes of the input. Using the pseudo-label of o_i as a ground truth, the whole network is trained to minimize the following classification objective:

$$\mathcal{L}_{cls} = \texttt{CrossEntropy}\left(y_p(o_i), \hat{y_{\mathcal{A}}}(a_i)\right) + \texttt{CrossEntropy}\left(y_p(o_i), \hat{y_{\mathcal{V}}}(v_i)\right), \tag{3.2}$$

where $y_p(o_i)$ represents the one-hot pseudo-label of o_i. Moreover, $\hat{y_{\mathcal{A}}}(a_i)$ 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. Consequently, after training the network with $\mathcal{L}_{cls}$, the AV-Net can predict pseudo-classes of audio and visual features by its auxiliary classifiers.

The AV-Net is trained by attention-learning objective (Eq. 3.1) and pseudo-label classification task (Eq. 3.2) in an alternate manner, as two objectives mutually improves the overall performance [28].

Chapter 4. Training the Video Inpainting Network

In this chapter, we describe a novel training methodology which uses the functionalities of the audio-visual network (AV-Net). The AV-Net described in Ch. 3 is adopted to support the video inpainting network (VI-Net) to capture the semantic cues from the audio signal. We provide details of two novel losses derived from the AV-Net – audio-visual attention loss (Sec. 4.1) and pseudo-class consistency loss (Sec. 4.2). Fig. 4.1 illustrates the overview of the proposed framework.

Throughout this chapter, we assume there is a pair of a corrupted video frame $\bar{v}$ and its ground truth frame v. Then, the VI-Net returns the inpainted frame $\hat{v}$, given the corrupted frame $\bar{v}$. Note that weights of the pretrained AV-Net are frozen while training the VI-Net.

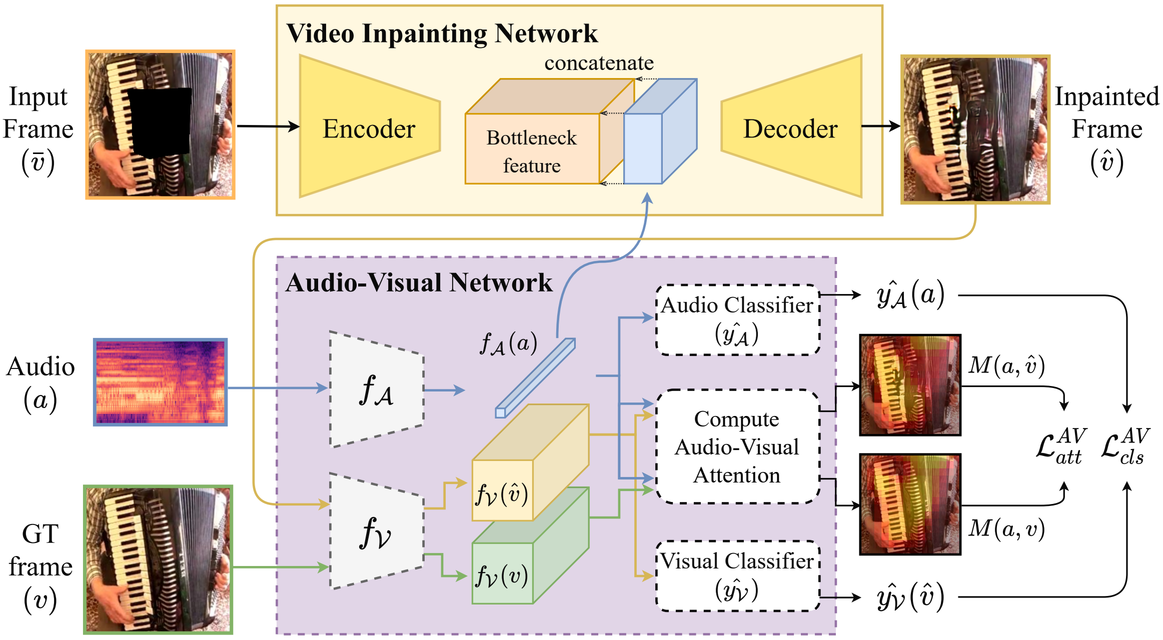

Figure 4.1: The overall schematic of training the **video inpainting network** (VI-Net) guided by the **audio-visual network** (AV-Net). The prior knowledge which the AV-Net has learned is transferred to the VI-Net through our **audio-visual attention loss** $\mathcal{L}^{AV}_{att}$ (Sec. 4.1) and **audio-visual pseudo-class consistency loss** $\mathcal{L}^{AV}_{cls}$ (Sec. 4.2). Both losses encourage the VI-Net to recover the frame whose content has high correspondence to the synchronized audio. **Audio-visual attention loss** is defined as a difference between the attention map of the inpainted frame $M(a,\hat{v})$ and that of the ground truth frame $M(a,v)$. **Audio-visual pseudo-class consistency loss** supports the visual contents to be class-consistent $\hat{y_\mathcal{V}}(\hat{v})$ to the pseudo-class of the concantenated audio feature $\hat{y_\mathcal{A}}(a)$. Components with dotted line indicates that their weights are frozen during training time.

4.1 Audio-Visual Attention Loss

We exploit the ability of 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 a highly responsive attention map $M(a, v)$ in the area matching with 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 contents of the audio-visual event are successfully recovered in $\hat{v}$, the attention maps $M(a, \hat{v})$ and $M(a, v)$ should be similar. 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 the two aforementioned attention maps would reduce the disparity between v and $\hat{v}$. Hence, we propose the following audio-visual attention loss:

$$\mathcal{L}_{att}^{AV} = \frac{1}{hw} \left\| M(a, v) - M(a, \hat{v}) \right\|_2^2 , \tag{4.1}$$

where h and w are height and width of $M(\cdot, \cdot)$, respectively. Eq. 4.1 indicates a mean squared error between two attention maps $M(a, v)$ and $M(a, \hat{v})$. Thus, this objective encourages the VI-Net to complete the corrupted frame in a way such that the audio-visual attention map $M(a, \hat{v})$ of the inpainted frame is similar to the attention map $M(a, v)$ of the ground truth frame. As a result, VI-Net can better restore the missing part of the sound-salient regions by matching their feature similarities to the synchronized audio. Hence, the inpainting network can better restore the missing part of sound-salient areas by filling it with contents or textures 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.

4.2 Audio-visual pseudo-class consistency loss

To further improve the performance of 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 can 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. 4.1). 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:

$$\mathcal{L}_{cls}^{AV} = \texttt{CrossEntropy}\left(\hat{y_{\mathcal{A}}}(a), \hat{y_{\mathcal{V}}}(\hat{v}) \right) , \tag{4.2}$$

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 already pretrained 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.

4.3 Total loss

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

$$\mathcal{L} = \lambda_{L_1}\mathcal{L}_{L_1} + \lambda_{adv}\mathcal{L}_{adv} + \lambda^{AV}_{att}\mathcal{L}^{AV}_{att} + \lambda^{AV}_{cls}\mathcal{L}^{AV}_{cls}. \tag{4.3}$$

$\mathcal{L}_{L_1}$ indicates the sum of L_1 losses computed for missing and valid region, which are individually normalized by the number of corresponding pixels. $\mathcal{L}_{adv}$ is the adversarial loss from Temporal PatchGAN [24], which enhances the spatio-temporal reality of synthesized video frames. Note that these two losses are borrowed from [22], which is our baseline VI-Net. The VI-Net is optimized jointly with our proposed audio-visual losses $\mathcal{L}^{AV}_{att}$ and $\mathcal{L}^{AV}_{cls}$. Hence, the network learns to minimize the audio-visual consistency as well as the visual difference. The weights for each loss are empirically set as follows: $\lambda_{L_1} = 1$, $\lambda_{adv} = 0.01$, $\lambda^{AV}_{att} = 2$, and $\lambda^{AV}_{cls} = 1$.

Chapter 5. Experimental Setup

5.1 Dataset and Preprocessing

Datasets. We evaluate the effectiveness of our method on two video datasets with audio-visual data pairs: **AVE** [11] and **MUSIC-Solo** [8] dataset.

- **AVE** dataset contains 4,113 video clips covering 29 categories of diverse real-life audio-visual events. Here, we examine whether our method enhances the performance on restoring videos of diverse events. We follow the official split of AVE dataset whose number of train/validation/test set is 3312/401/402.
- **MUSIC-Solo** dataset contains 493 video clips with 11 categories that exclusively cover solo performances of different musical instruments. This dataset is employed to evaluate the transferability of our methods on narrower audio-visual contexts. We randomly split MUSIC-Solo dataset into 343/50/100 for each train/validation/test set since there is no designated split.

Moreover, we evaluate our methods on two types of maskings: **I-mask** and **S-mask**.

- **I-mask** irregularly blinds the pixels with random strokes and shapes. We adopt the subset of NVIDIA Irregular Mask Dataset [29]. For testing, we randomly pick three I-masks whose blinding ratio is 20.0%, 27.7%, and 28.4%, each. Note that I-masks are adopted for evaluating the general performance of the video inpainting network.
- **S-mask** is designed to blind the regions which corresponds to the sounding object. Thus, with S-masks, we evaluate whether the video inpainting network can recover these salient regions that are partially deteriorated. We collect S-masks by eroding[1] the object mask m_i mentioned in Sec. 3.2 until the spatial area of the masking covers about 20% of the image. This ratio refers to the approximate proportion of the region that the sounding object occupies in the video frame.

Preprocessing. Given a video clip of arbitrary length, 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 synchronized audio is sampled by retrieving a 1-second segment whose midpoint of the segment corresponds to the given frame. The audio segment is converted to the log-scale mel spectrogram[2] 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.

---

[1] We implement this process by using `erode` method in `opencv-python` package.

[2] We implement this process by using `melspectrogram` method in `librosa` package.

5.2 Implementation Details

The Audio-Visual Network. We follow [3, 27] 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 [27]. In detail, we change the stride of the first convolutional layer in the last residual block from 2 to 1 for getting a feature map with a larger spatial size of 14×14. Furthermore, the input channel of the first convolutional layer in the audio sub-network is modified to 1 for processing the input of a single-channel spectrogram. We also note that both sub-networks are initialized with ImageNet-pretrained weights. As both sub-networks are simple variants of ResNet-18, the channel dimension of their output is 512. We also note that both sub-networks are initialized with ImageNet-pretrained weights. Thus, the visual sub-network produces the feature map of $\mathbb{R}^{14\times14\times512}$, while the audio sub-network gives the vector of $\mathbb{R}^{512}$ from a pooled feature map. To compute the attention between features from two modalities, we compress the channel dimension of the visual feature map by 1×1 convolution, resulting in $\mathbb{R}^{14\times14\times128}$. Similarly, we use fully connected layers for projecting the audio feature vector into 128-dimension. The audio-visual attention map is finally computed between these two compressed representations.

Visual and audio classifiers in the AV-Net are individually implemented by a fully connected layer which produces a logit vector by getting the original representation of each modality. Note that these linear classifiers are re-initialized to random weights every time after K-means clustering is conducted.

Video inpainting baseline. We adopt one of the state-of-the-art architectures, the Spatial-Temporal Transformer Network (STTN) [22] as our baseline. The major component of this network is a transformer block that computes the similarity between all of the spatio-temporal patches within the encoded frame feature. This ensures the spatio-temporal consistency of the inpainted contents by guiding the network when and where to attend among consecutive input frames.

As our major interest lies in inpainting videos with audio-visual events, our choice of video dataset is different from the original work [22]. Therefore, we obtain our baseline by training the STTN with the aforementioned datasets from scratch without the audio signals. We fully reproduce the identical architecture, referring to its official implementation[3].

Training details. To train the AV-Net, we adopt Adam optimizer with the learning rate of `5e-5` for training the network with AVE dataset and `1e-4` for MUSIC-Solo dataset. The batch size is set to 32 for both datasets. Furthermore, we set the threshold value of 0.07 to obtain the binary mask while collecting object representations. The number of clusters is set to 10 for extracting pseudo labels from K-means clustering. While training the AV-Net for 4 epochs total, the learning rate is decayed by 0.1 after 2 epochs. On the other hand, for training VI-Net, the weights of pretrained AV-Net are frozen in training time. In the experiments with AVE dataset, we train the VI-Net from randomly initialized weights. For the 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. Furthermore, for MUSIC-Solo dataset, due to its lacking of training data, we fine-tune the VI-Net pretrained on the 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. The batch size is set to 8 for both datasets.

---

[3] The official source code of the STTN in GitHub: https://github.com/researchmm/STTN

Chapter 6. Result

6.1 Evaluation Metrics

The quantitative result is reported using three widely-used metrics: PSNR [18], SSIM [30], and video-based Fréchet Inception Distance (VFID) [24]. In detail, PSNR and SSIM are standardized metrics for assessing the quality of visually synthesized material. VFID has been recently employed to quantify the perceptual distance between two sets of video features. To compute VFID, video features are extracted using a pretrained I3D network [31] to compute VFID following [22, 24].

6.2 Analysis and Discussion

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

Table 6.1: 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		
	Audio	$\mathcal{L}_{att}^{AV}$	$\mathcal{L}_{cls}^{AV}$	PSNR↑	SSIM↑	VFID↓	PSNR↑	SSIM↑	VFID↓
Baseline	✗	✗	✗	29.49	93.85	4.316	26.47	91.88	5.706
+ Ours	✓	✗	✓	29.60	93.84	4.191	26.95	92.38	5.205
+ Ours	✓	✓	✗	29.66	**93.87**	4.221	27.05	92.40	5.194
+ Ours	✓	✓	✓	**29.68**	93.84	**4.092**	**27.12**	**92.53**	**4.929**

Table 6.2: 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.

We test our method on 4 different experimental setups derived from the combinations of video and mask datasets mentioned in Sec. 5.1. Table 6.1 shows that adopting our proposed audio-visual objectives outperforms the visual-only baseline on AVE dataset for all suggested metrics. As shown in Table 6.2, our method also performs substantially well on the MUSIC-Solo dataset with video scenes strictly related to musical instruments. On the MUSIC-Solo dataset with I-masks, our methods show very similar PSNR and SSIM scores compared to the baseline. However, under the same settings, our method, especially the one using both $\mathcal{L}_{att}^{AV}$ and $\mathcal{L}_{cls}^{AV}$, shows significantly lower VFID scores. In other words, although our approach shows comparable results in terms of pixel-level difference, we observe a great improvement in terms of VFID which implies that ours can further improve the perceptual reality. Performance improvements over the 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 6.1 and 6.2 imply that our two losses harmoniously give a positive impact on the inpainting quality, with the audio-visual attention loss showing higher competitiveness.

One interesting point is that performance gains on S-masks are more significant than those on I-masks. As shown in Table 6.1, on the AVE dataset masked with I-masks, our method of applying both two losses improves the baseline PSNR and VFID by 0.42 and 0.365, respectively. On the same dataset with S-masks, our method shows larger PSNR improvement of 0.74, and 0.592 in case of VFID. The same tendency is shown in Table 6.2 on the MUSIC-Solo dataset. In the case of I-masks, PSNR and VFID improvement shows each 0.19 and 0.224 compared to the baseline. On the other hand, improvements are greater in the case of S-masks, showing PSNR improvement of 0.65 and 0.777 in terms of VFID metrics. 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 network to better restore regions corresponding to the audio-visual events.

Fig. 6.1 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 more 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 contents, while preserving the audio-visual coherency.

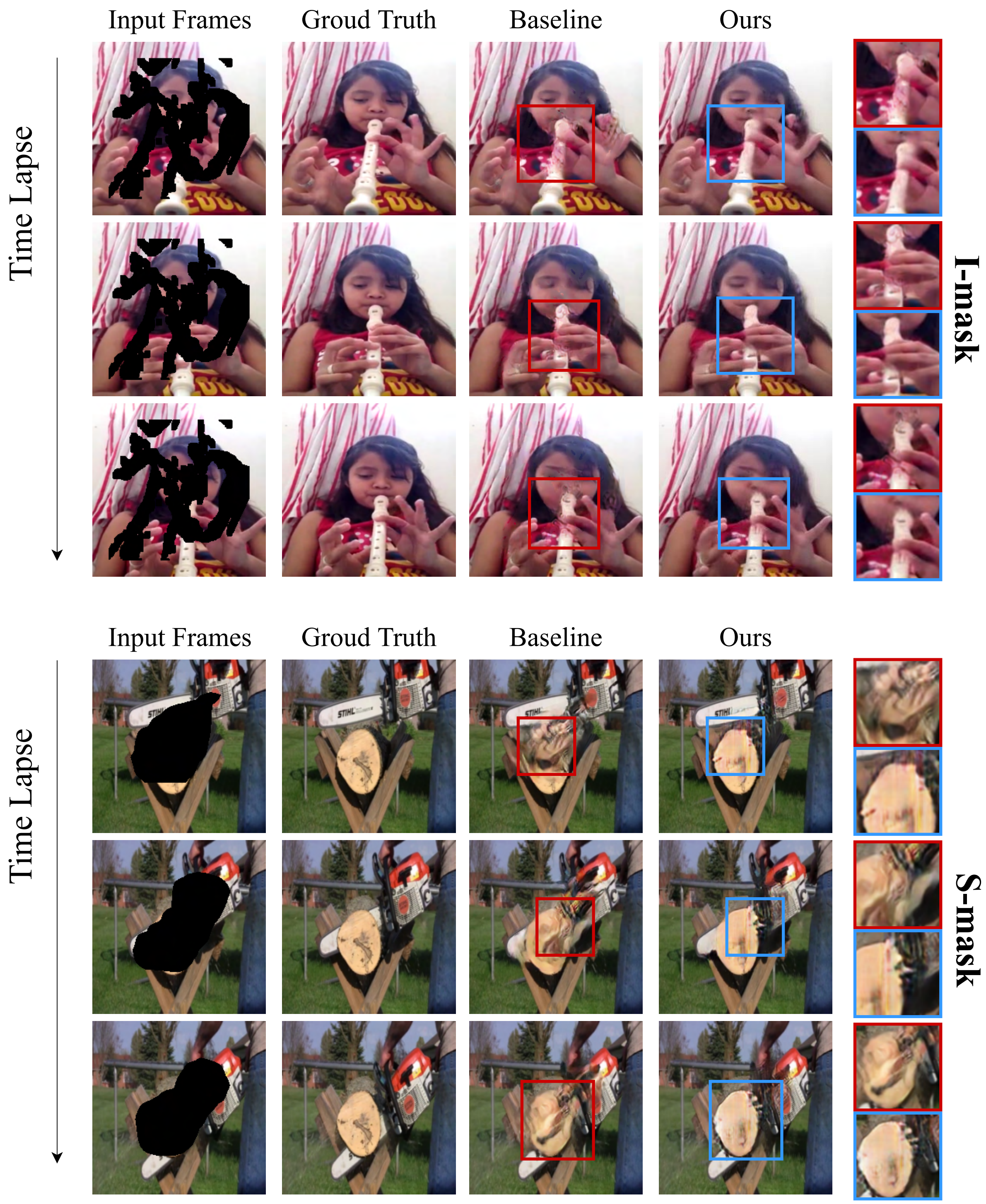

Figure 6.1: Qualitative results of two samples from the AVE dataset blinded by **I-masks** (top) and **S-masks** (bottom). While the baseline STTN shows more artifacts around the sounding object and produces blurry contents, our method produces more realistic and clearer results with less artifacts.

Chapter 7. Conclusion

In this paper, we investigate a novel approach to using the 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, securing the correlation and classification capability. These capabilities guide the video inpainting network to grab extra information from the synchronized audio via two novel losses we propose – 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 inpainting performance of the video inpainting network compared to the baseline.

Since our work manages scenarios with a single sound source, future work can take a comprehensive approach to complicated audio-visual scenes, such as scenes of multiple sounding objects. We believe that our work paves the new avenue of using audio for other visual restoration tasks such as video super-resolution and colorization.

Bibliography

[1] Kyuyeon Kim, Junsik Jung, Woo Jae Kim, and Sung-Eui Yoon, "Deep video inpainting guided by audio-visual self-supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022 (under review).

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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 (TOG)*, vol. 37, no. 4, 2018.

[10] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba, "Ambient sound provides supervision for visual learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 801–816.

[11] 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.

[12] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu, “Deep cross-modal audio-visual generation,” in *Proceedings of the on Thematic Workshops of ACM Multimedia*, 2017, pp. 349–357.

[13] Wangli Hao, Zhaoxiang Zhang, and He Guan, “Cmcgan: A uniform framework for cross-modal visual-audio mutual generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018, vol. 32.

[14] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang, “Vision-infused deep audio inpainting,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 283–292.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Andrew Owens and Alexei A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[27] 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.

[28] 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.

[29] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro, "Nvidia irregular mask dataset," in *https://nv-adlr.github.io/publication/partialconv-inpainting*, 2018.

[30] 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.

[31] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.

Acknowledgments in Korean

지난 2년 간의 시간을 돌이켜보면 감사함을 전하고 싶은 분들이 참 많습니다. 먼저 연구에 대한 혜안을 주시고 아낌없는 지원과 격려를 해주신 윤성의 교수님께 깊은 감사의 말씀을 전합니다. 교수님의 지도로 연구 과정에서 용기를 잃지 않고, 차근차근 앞으로 나아갈 수 있었습니다.

대학원 생활을 함께한 저희 연구실 식구 분들께도 감사의 말씀을 전합니다. 사소한 질문에도 친절하게 답해주시고 연구실 생활에 적응할 수 있도록 많은 도움을 주신 연구실 선배님들께 감사드립니다. 가끔 담소를 나누며 저에게 힘을 주셨던 연구실 동기 분들과 후배님들께도 감사의 말씀을 전합니다. 항상 배울점만 가득했던 크고 작은 선생님들인 연구실 선후배님들을 곁에 둘 수 있었던 것은 제 대학원 생활의 크나큰 행운이었습니다.

제 부족했던 연구를 무사히 논문으로 승화시키는 데에 많은 힘과 시간을 투자해주신 정준식 형님과 김우재 후배님께 감사의 말씀을 전합니다. 위클리 미팅을 함께하며 제 연구 내용을 경청해주시고 값진 조언을 아끼지 않으신 비전 팀의 김태영 선배님, 임우빈 선배님, 제 옆자리를 지켜주신 김재윤 선배님, 윤욱 (Xu Yin) 선배님, 안국원 (Guoyuan An) 선배님, 조윤기 선배님 모두 진심으로 감사드립니다. 또한, 연구실의 든든한 버팀목이 되어주신 권용선 선배님, 강민철 선배님, 안인규 선배님, 신희찬 선배님, 장충수 선배님, 김민철 선배님, 제 맞은편 자리를 지켜주신 김재윤 선배님께 감사의 말씀을 전합니다. 2년의 기간을 함께한 연구실 동기인 형열님, 민성님, 세빈님께도 감사드립니다. 길고도 짧은 기간 동안 함께한 김진원 님, Minh-Tuan Tran 님, 한규범 님, Pei Wang 님께도 감사의 말씀을 전합니다. 대학원을 졸업하시고 먼저 사회로 진출하신 김수민 박사님, 박훈민 선배님, 장진혁 선배님, 조인영 선배님, 조창호 선배님께도 감사의 말씀을 전하고 싶습니다. 연구에 집중할 수 있도록 행정적 지원을 아끼지 않으신 김슬기나 선생님께도 감사의 말씀을 전합니다.

긴 시간 동안 동고동락해오며, 항상 저에게 힘이 되어준 연재에게 고마운 마음을 전합니다. 항상 웃음을 잃지 않도록 함께한 고등학교 친구들에게도 고마움을 전합니다. 마지막으로, 제 모든 것의 원동력이 되어주시고 저를 믿고 지켜봐주신 부모님과 누나에게도 감사함과 사랑의 마음을 전합니다. 항상 겸손하고, 타인에게 베풀며 힘이 될 수 있는 지혜로운 사람이 되겠습니다. 다시 한 번 감사드립니다.

김규연 올림.

Curriculum Vitae in Korean

이름: 김규연

생년월일: 1996년 7월 15일

전자주소: kyuyeonpooh@gmail.com

학력

2012. 3. – 2015. 2.	중동고등학교
2016. 3. – 2020. 2.	성균관대학교 소프트웨어학과 (학사)
2020. 3. – 2022. 2.	한국과학기술원 전산학부 (석사)

경력

2019. 9. – 2019. 12.	호주 연방과학산업연구기구 (CSIRO Data61) 인턴
2020. 9. – 2020. 12.	한국과학기술원 전산학부 조교 (CS206, 데이타구조)
2021. 3. – 2021. 6.	한국과학기술원 전산학부 조교 (CS492, 전산학특강<데이터 사이언스 개론>)

연구업적

1. **Kyuyeon Kim**, Junsik Jung, Woo Jae Kim, and Sung-Eui Yoon, *Deep Video Inpainting Guided by Audio-Visual Self-Supervision*, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2022 (under review).
2. **Kyuyeon Kim** and Sung-Eui Yoon, *Image-Bridged Multimodal Half-Supervised Learning*, KIISE Transactions on Computing Practices (KTCP), 2021.
3. Yansong Gao, Minki Kim, Sharif Abuadbba, Yeonjae Kim, Chandra Thapa, **Kyuyeon Kim**, Seyit A. Camtepe, Hyoungshick Kim, and Surya Nepal, *End-to-End Evaluation of Federated Learning and Split Learning for Internet of Things*, International Symposium on Reliable Distributed Systems (SRDS), 2020.
4. Sharif Abuadbba, **Kyuyeon Kim**, Minki Kim, Chandra Thapa, Seyit A. Camtepe, Yansong Gao, Hyoungshick Kim, and Surya Nepal, *Can We Use Split Learning on 1D CNN Models for Privacy Preserving Training?*, ACM Asia Conference on Computer and Communications Security (ASIACCS), 2020.