THE IMPACT OF SPATIO-TEMPORAL AUGMENTATIONS ON SELF-SUPERVISED AUDIOVISUAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning of auditory and visual perception has been extremely successful when investigated individually. However, there are still major questions on how we could integrate principles learned from both domains to attain effective audiovisual representations. In this paper, we present a contrastive framework to learn audiovisual representations from unlabeled videos. The type and strength of augmentations utilized during self-supervised pre-training play a crucial role for contrastive frameworks to work sufficiently. Hence, we extensively investigate composition of temporal augmentations suitable for learning audiovisual representations; we find lossy spatio-temporal transformations that do not corrupt the temporal coherency of videos are the most effective. Furthermore, we show that the effectiveness of these transformations scales with higher temporal resolution and stronger transformation intensity. Compared to self-supervised models pre-trained on only sampling-based temporal augmentation, self-supervised models pre-trained with our temporal augmentations lead to approximately **6.5%** gain on linear classifier performance on AVE dataset. Lastly, we show that despite their simplicity, our proposed transformations work well across self-supervised learning frameworks (SimSiam, MoCoV3, etc), and benchmark audiovisual dataset (AVE).

1 INTRODUCTION

Visual and auditory perception are the two most utilized human sensory systems in our day-to-day lives. The integration between the two systems allow us to capture rich representations of the environment around us. For instance, in an action recognition task, without sound it can be hard to tell whether a child is babbling, laughing, or coughing using only video frames. However with sound integrated, the problem becomes substantially easier to solve. Applications of learning efficient audiovisual representations can range from audiovisual localization (Tian et al., 2018b; Arandjelović & Zisserman, 2018; Senocak et al., 2018) and separation (Zhao et al., 2018; Gao et al., 2018; Zhao et al., 2019) to action Recognition (Kazakos et al., 2019; Cartas et al., 2019; Xiao et al., 2020) and speech recognition (Nagrani et al., 2017; 2018).

Recently, contrastive self-supervised learning is at the forefront in learning abstract representations from unlabeled visual or auditory data (He et al., 2020; Chen et al., 2020a; Al-Tahan & Mohsenzadeh, 2021; Ye et al., 2019; Grill et al., 2020). One crucial component that allows these contrastive frameworks to excel at learning representations is the augmentations used during self-supervised pre-training. Hence, extensive augmentation search has been investigated for images (Chen et al., 2020a), audio (Al-Tahan & Mohsenzadeh, 2021), and videos (Feichtenhofer et al., 2021b; Qian et al., 2021). Previously proposed video augmentations exploited the temporal dimension of videos, defined as either changing the number of positive clips supplied to the self-supervised loss (Feichtenhofer et al., 2021b) or the sampling method of positive clips (Qian et al., 2021). Although, these temporal augmentations were shown to aid in learning better spatio-temporal representations, temporal augmentations that directly change spatial information temporally (spatio-temporal augmentations) while preserving video coherency is yet to be investigated.

In this work, we investigate audiovisual integration in a contrastive self-supervised setting to learn efficient representations. In order to accomplish this, we introduce four major components that are important to nourish the learning of spatio-temporal representations. We:

- Demonstrate a simple pipeline to learning efficient audiovisual representations that can be adopted with various existing contrastive frameworks (i.e. SimCLR, MoCo ...).
- Introduce four spatio-temporal augmentations that allow models to generalize better to the downstream tasks, compared to control models trained without the proposed spatio-temporal augmentations.
- Extensively study hyper-parameters of the spatio-temporal augmentations and verify the effectiveness of the augmentations as we scale the temporal resolutions of videos.
- Use the proposed augmentations and investigate temporally aligning augmentations for audiovisual integration.

2 Related Works

2.1 Self-Supervised Learning

2.1.1 Images

Self-supervised learning on images has been studied extensively over the last few years, serving as an essential benchmark. Generally self-supervised learning methods exploit the inherit structure of the training data to derive a supervisory signal, called the *pre-text* task. By training on the pre-text task, the aim is to derive effective visual representations from unlabel images, which can be used for various downstream tasks (e.g. classification, segmentation, ...). Some early pre-text tasks leveraged context (Pathak et al., 2016), jigsaw puzzle (Noroozi & Favaro, 2017), image rotation (Gidaris et al., 2018), relative patch spatial location (Doersch et al., 2016), and various other image structural characteristics (Zhang et al., 2016; 2017; Larsson et al., 2017; Bojanowski & Joulin, 2017). Numerous pre-text tasks even utilized video frames to learn efficient image representations (Wang & Gupta, 2015; Vondrick et al., 2018; Pathak et al., 2017; Gordon et al., 2020; Purushwalkam & Gupta, 2020; Jayaraman & Grauman, 2016). More recently, *contrastive learning* pre-text tasks has been widely adopted in learning efficient visual representations (He et al., 2020; Chen et al., 2020a; Ye et al., 2019; Grill et al., 2020; Tian et al., 2020; Hénaff et al., 2020; van den Oord et al., 2019; Chen & He, 2020). In short, the learning objective of contrastive learning is to maintain consistent representations between augmented views originating from the same image, while maximizing the representations between views from different images.

2.1.2 Audio

Similar to images, auditory data also has been rapidly progressing towards self-supervised auditory representation learning (van den Oord et al., 2019; Wang & van den Oord, 2021; Al-Tahan & Mohsenzadeh, 2021; Baevski et al., 2020a;b). Prior works derived efficient auditory representations using videos by predicting whether the visual and audio signals come from the same video (Aytar et al., 2016; Arandjelović & Zisserman, 2017; 2018; Korbar et al., 2018; Owens & Efros, 2018; Alwassel et al., 2020a; Alayrac et al., 2020; Patrick et al., 2020). As discussed earlier, contrastive learning relies heavily on augmented views to construct representations to become less sensitive to sensory-level invariance. Wang & van den Oord (2021) constructed those augmented views for auditory data by contrasting between raw-audio and audio frequency features (e.g. Short-time Fourier transform) using two distinct different models. Alternatively, Al-Tahan & Mohsenzadeh (2021) reduced the reliance on raw-audio by investigating six transformations that specifically tackle auditory data, while only using one model during training. Similar to visual representation learning (Chen et al., 2020a), some augmentations like pitch shift and fade in/out were shown to result in more efficient representations (Al-Tahan & Mohsenzadeh, 2021).

2.1.3 Audiovisual

Audiovisual integration using self-supervised learning by exploiting audiovisual correspondence has been extensively explored (Arandjelović & Zisserman, 2017; Aytar et al., 2016; Owens & Efros,

2018; Owens et al., 2016; Korbar et al., 2018; Hu et al., 2019; Alwassel et al., 2020b). The benefit of learning efficient audiovisual representations can aid wide range of tasks beyond video recognition (Nagrani et al., 2018; Arandjelović & Zisserman, 2018; Kazakos et al., 2019; Zhao et al., 2018; Cartas et al., 2019; Zhao et al., 2019). Qian et al. (2021) investigated the spatio-temporal component of video frames using contrastive learning for learning video representations. Qian et al. (2021) found that maintaining temporal consistency in-regard to spatial augmentation is crucial for better representations. Furthermore, they found that the temporal sampling strategy for the contrasted positive clips is also important in learning efficient representations. Alternatively, Feichtenhofer et al. (2021b) defined temporal augmentations as the number of clips sampled at different temporal locations as positive samples and found that by increasing the number of clips (≥ 2), we can obtain better representations.

In this work, we adopt findings from mono-domain contrastive learning frameworks on audio and video domains. In particular, for the video domain, we utilize the sampling strategy with a monotonically decreasing distribution (Qian et al., 2021). Furthermore, we maintain consistent spatial augmentations across frames within each clip. For the audio domain, we utilized the auditory augmentations that affect the frequency structure of the signal rather than the temporal component of a signal such as pitch shift (Al-Tahan & Mohsenzadeh, 2021). Lastly, to nourish better audiovisual representations, we introduce four spatio-temporal augmentations and extensively study the effectiveness of these augmentations on generalizability to downstream tasks.

3 METHODS

3.1 CONTRASTIVE LEARNING FRAMEWORKS

Contrastive Learning frameworks generally aims to maximize similarity between representations of the same samples that are augmented differently $\tilde{x}$ (positive samples) and minimize similarity between representations of different samples (negative samples). In this paper, our positive samples are the set of clips extracted from each video and our contrastive loss in principle follows the InfoNCE objective (Oord et al., 2018; Chen et al., 2020a):

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_k\right)/\tau\right)} \tag{1}$$

where $\mathcal{L} = \sum_{i,j}^{N} \mathcal{L}_{i,j}$, $\mathbf{1}_{[k\neq i]} \in 0, 1$ is an indicator function evaluating to 1 iff $k \neq i$, τ denotes a temperature parameter, and $\mathbf{z}$ denotes ℓ_2 normalized encoded representations of a given augmented clip. N is the number of training samples within a mini-batch, (i, j) are positive clips from each video. The loss is computed across all positive clips, in a mini-batch. $\text{sim}\left(\mathbf{u}, \mathbf{v}\right) = \mathbf{u}^T\mathbf{v}/||\mathbf{u}||||\mathbf{v}||$ denotes the cosine similarity between two vectors $\mathbf{u}$ and $\mathbf{v}$.

In this paper, we investigate multitude of frameworks for audiovisual representations learning that utilizes a variant of the InfoNCE objective:

1. **SimCLR** (Chen et al., 2020a) uses the same objective as Eq.1, where the representations of different clips within a mini-batch are treated as the negative samples. SimCLR adopt the encoder to extract the representations of different clips to compute the contrastive loss with the gradient of both views flowing through the encoder. Hence, this method relies heavily on negative samples to prevent collapsing.
2. **MoCo** (Chen et al., 2021) replaces the second identical encoder with a momentum encoder, θ_m. The momentum encoder parameters are a moving average of the encoder θ and updated at each step: $\theta_m \leftarrow m\theta_m + (1-m)\theta$ where $m \in (0, 1)$ is the hyper-parameter that dictates the degree of change. There is no gradient flowing through the momentum encoder. Our implementation follows Chen et al. (2021) design choices rather than Chen et al. (2020b); this means that we replace shuffling batch normalization (BN) with sync BN, the projection head is a 3-layer MLP, and the prediction head is a 2-layer MLP. The prediction head is stacked on top of the projection head, however, the prediction head does not stack on top of momentum encoder representations. Lastly, we preserve the memory queue, as we utilize relatively small batch size.

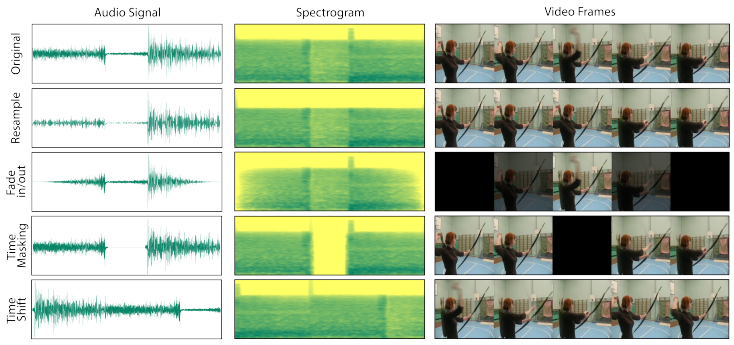

Figure 1: Temporal Augmentations. Each row demonstrate the effect of an augmentation on the raw waveform, mel-spectrogram, and video frames, respectively. The first row, represent the original video with its respective audio without any transformation applied.

3. **BYOL** (Grill et al., 2020) can be viewed as a variant of MoCo that does not use negative samples. Hence, the loss function consists of only the numerator part of Eq. 1.
4. **SimSiam** (Chen & He, 2020) similar to SimCLR, this method uses identical second encoder. However, SimSiam does not use negative samples and the gradient only flow through one of the encoders. SimSiam can be thought of as BYOL without the momentum encoder.

3.2 Audiovisual Encoder

For each clip, we encode the video frames and audio spectrograms using a ResNet (He et al., 2015) variant. Unless mentioned otherwise, we utilized ResNet-18 for both streams. For the video encoder, we follow the design of SlowFast network with the modifications proposed in CVRL (Feichtenhofer et al., 2019; Qian et al., 2021). For the audio encoder, we followed the design of (Al-Tahan & Mohsenzadeh, 2021; Kazakos et al., 2021), however due to memory restrains we apply max-pooling to the temporal dimension, contrary to the implementation proposed by Kazakos et al. (2021). All models were trained from random initialization with 4 and 8 NVIDIA v100 Tesla GPUs.

3.3 Data Augmentations

The choice of augmentations are crucial component of contrastive learning to construct view invariant in which makes the model learn efficient representations and become less sensitive to variations in the input data (Chen et al., 2020a; Al-Tahan & Mohsenzadeh, 2021; Gidaris et al., 2018; Noroozi & Favaro, 2017). For images, Chen et al. (2020a) have shown that the combination of random cropping followed by resize back to the original size, random horizontal flip, random color distortions, and random Gaussian blur is crucial to achieve a good performance. For sounds, Al-Tahan & Mohsenzadeh (2021) have shown that the combination of frequency and temporal guided augmentations yield the best performance (i.e. pitch shifting and time masking). We incorporate augmentations from both domains in our framework, however for transformations that operate on the temporal axis, we apply them for both auditory and visual data (see Section 3.3.1). For training dataset, we utilize AVE dataset (Tian et al., 2018a) because audios and videos correspond, in the sense that the sound source is always visually evident within the video clip.

3.3.1 Temporal Augmentations

Figure 1 shows the effect of temporal augmentations on raw waveform, mel-spectrogram, and video frames, respectively. For each augmentation we defined parameter $\alpha \in [0, 1]$ which controls the maximum intensity of the augmentations across the temporal dimension; where 0 in-tales that none of the video is augmented and 1 means that the augmentation affect all the video in the temporal dimension. In the current section, we will describe the specifics of each augmentation:

1. **Fade in/out (FD)**: Gradually increases/decreases the intensity of the audio signal or video frames starting from the beginning and end of the video to the middle of the video. The

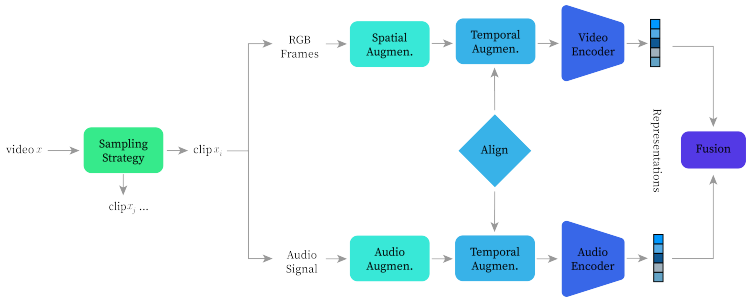

Figure 2: Illustration of self-supervised pre-training pipeline. Given a video x, we sample 2 clips using monotonically decreasing probability distribution, followed by augmenting each stream with domain specific augmentations and temporal augmentations.

degree of the fade is either linear, logarithmic, exponential, quarter sine, or half sine (applied with uniform probability distribution). The size of the fade for either side of the video is another random parameter applied using a uniform probability distribution (each side of the fade is independent). For instance, $\alpha = 0.5$ for one side means that $1/4$ of the total video clip is used for the fade effect.

2. **Time Masking (TM)**: Randomly masks a small segment of the audio signal or video frames with normal noise or a constant value. We randomly selected the location of the masked segment and the size of the segment from a uniform distribution. α controls for the maximum size of the masked segment.
3. **Time Shift (TS)**: randomly shifts the audio signal or video frames forwards or backwards temporally. Samples/frames that roll beyond the last position are re-introduced at the first position (rollover). The degree and direction of the shifts were randomly selected for each audio. The maximum degree that could be shifted was based on α, while, the minimum was when no shift applied.
4. **Resample (RE)**: Randomly apply temporal down-sample, followed by an up-sample back to the original shape. The RE transformation results in lower resolution audio and video signals with repetitive samples. For instance, $\alpha = 0.5$ for an audio signal with 44.1 kHz sampling rate would be down-sampled to 22 kHz and up-sampled back to 44.1 kHz. Further for video frames with 8 fps, $\alpha = 0.5$ would down-sample the video to 4 fps and up-sample back 8 fps.

3.3.2 Temporal Augmentation Alignment

Augmentations that maintain the spatio-temporal coherency within videos have been shown to yield better representations (Qian et al., 2021). We incorporate an analogous paradigm by aligning augmentations temporally across audio and video streams. For instance, with fade in/out augmentation, we align the location of the fade (beginning or end) and strength of the augmentation by making them consistent across the two streams. Similar to the fade in/out augmentation, time masking and time shift follow the same standard; where the strength and the temporal location of the augmentation is made to be consistent across streams. Lastly, for the resample augmentation, we simply maintain consistency just for the strength (α) of the augmentation.

3.4 Evaluation

To evaluate the efficiency of learned representations in our experiments, we follow common practice in self-supervised learning literature (Chen et al., 2020a; He et al., 2020). Specifically, after training the encoder on the contrastive learning pre-text task, we fix the weights in the encoder and train a linear classifier on top of it to evaluate the efficiency of the encoder learned representations. During the training of the linear classifier, we only train using domain specific augmentations, without the temporal augmentations. This measure was done to maintain consistent pipeline when comparing encoders trained using different temporal augmentations (i.e. 4.2). The linear classifier training uses base lr = 30 with a cosine decay scheduler for 100 epochs, weight decay = 0, momentum = 0.9, batch

size = 256 and SGD optimizer, following Chen & He (2020); He et al. (2020). During testing, we follow common practice of uniformly sampling 10 clips from each video with a 3-crop evaluation (Feichtenhofer et al., 2019; Wang et al., 2018) where each clip has spatial shape of 256×256. The final prediction is the averaged softmax scores of all clips.

3.5 Implementation Detail

Following Chen & He (2020), all experiments were trained using SGD optimizer with a learning rate of $lr\times$ Batch Size / 256 ($lr = 0.1$), momentum of 0.9, and weight decay of 10^{-4}. Weight decay were not applied to bias parameters and batch normalization layers. Furthermore, batch normalization layers were synchronized across devices. For learning rate, we used linear warm-up for the first 10 epochs (Goyal et al., 2018), and cosine decay schedule without restarts (Loshchilov & Hutter, 2017). Unless specified, we used batch size of 128 for all experiments. Consistent with (Qian et al., 2021), each clip is 1.28 seconds long and sampled the two positive clips using the same strategy. That is, given an input video T, we draw a time interval t from a monotonically decreasing distribution over $[0, T]$. Following that, we uniformly sample the first clip interval from $[0, T - t]$, while the second clip is delayed by t after the first.

4 Results

4.1 Auditory Augmentations

Auditory data can be augmented in both temporal and frequency domain (Al-Tahan & Mohsenzadeh, 2021). For the audio stream, we initially adopt frequency augmentations followed by temporal augmentations. Al-Tahan & Mohsenzadeh (2021) have shown colored noise and pitch shifting as viable options. Table1 investigates the effect of different auditory augmentations and the location of applying the auditory augmentations relative to the temporal augmentations on linear classifier top-1/5 accuracy after being pre-trained (with contrastive loss) on AVE (Tian et al., 2018a) dataset for 500 epochs. Results suggest that pitch shifting on average shows **+5.9%** higher performance compared to colored noise. This result is consistent with (Al-Tahan & Mohsenzadeh, 2021). Furthermore, we observe that placing the auditory augmentations either before/after the spatio-temporal augmentations can yield $\sim$**1%** difference in performance. This can also depend on the auditory augmentation applied, for instance, pitch shifting benefit when placed before the spatio-temporal augmentations, while colored noise benefit when placed after. To stay consistent, moving forward we will only use pitch shifting before the spatio-temporal augmentations.

Table 1: The impact of type and location of audio augmentations applied before temporal augmentations. RE was applied for pre-training all models. Top-1/5 accuracy of linear classifiers pre-trained on AVE dataset for 500 epochs.

Audio Augmentation	Location			
	Before		After	
	Accuracy (%)		Accuracy (%)	
	top-1	top-5	top-1	top-5
Colored Noise	58.48	90.55	59.95	92.04
Pitch Shift	**65.42**	**93.28**	64.18	91.29

4.2 Temporal Augmentations

Most recent contrastive learning frameworks relies heavily on augmentations to derive the contrastive loss. While previous video self-supervised frameworks investigated the effect of temporal sampling of clips on learnt representations (Qian et al., 2021), a wider range of temporal augmentations that can directly effect the learned representations has not been explored. In this section, we vary the strength, sequence, and across domain application of the proposed spatio-temporal augmentations to explore their effect on predictive performance:

4.2.1 STRENGTH

Before exploring multiple spatio-temporal augmentations in sequence, we first investigate the effect of changing the strength of each temporal augmentation during the training on the contrastive pre-task. Table 2 shows the effect of varying α for each of the spatio-temporal augmentations. As described in Section 3.3.1, α controls the maximum intensity of the augmentations across the temporal dimension. Similar to prior work using spatial augmentations (Chen et al., 2020a; Feichtenhofer et al., 2021a), we observe that stronger augmentations (higher α) generally increase performance of the linear classifier. With the exception being Time Shift, where performance does not significantly change across different α. The best performing augmentation is Resample with **+5.0%** higher performance than the second best performing temporal augmentation. Furthermore, Resample shows the most performance difference when changing α, with **+6.5%** difference going from low to high α.

Table 2: The impact of α variable for each of the augmentations. Top-1 accuracy of linear classifiers pre-trained on AVE dataset for 500 epochs.

Augmentation	Top-1 Accuracy (%)		
	$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 0.75$
Resample	58.96	62.69	**65.42**
Fade in/out	49.75	**58.71**	57.21
Time Masking	54.73	57.71	**60.45**
Time Shifting	**57.96**	56.72	57.71

4.2.2 SEQUENCE

To systematically investigate the impact of temporal augmentations and their sequential ordering, we explore a composition of the proposed augmentations during pre-training on the contrastive task. Figure 3 shows linear classifier top-1 performance on AVE dataset. The diagonal line represents the performance of single augmentation, while other entries represent the performance of paired augmentations. Each row indicates the first augmentation and each column shows the second augmentation applied sequentially. The last column depicts the average when the augmentation was applied first, while the last row shows the average performance over the corresponding augmentations when they were applied the second. The bottom right element serve as a control where no spatio-temporal augmentation was applied (**58.7%**). We observe that Resample plays a critical role in learning good representation. As just by including the Resample transformations during contrastive pre-training, we at least attain **61.7%**, which outperforms all models that are pre-trained on other spatio-temporal augmentations.

4.2.3 ACROSS DOMAIN APPLICATION

To investigate the impact of these spatio-temporal augmentations on the learnt representations of multiple domains, we restrain the temporal augmentations to one domain and explore across domains temporal alignment. In Table 3, we select the two top competing augmentations from Section 4.2.2: Resample only (**65.4%**) and Resample + Time Shift (**64.7%**). Without spatio-temporal augmentations, the linear classifier top-1 accuracy is **58.7%**. We observe that on average compared to no spatio-temporal augmentations, applying these augmentations only to the audio stream produces better representations compared to video stream (**5.23%** vs **2.12%**; respectively). We further tested whether aligning the intensity (α) and location (does not apply for Resample) of the augmentations temporally would result in better representations. However, we observe that on average not aligning the augmentations temporally produces better representations (**+2.37%** improvement).

We verify the effectiveness of the proposed spatio-temporal augmentations by varying the temporal resolution of both video and audio streams independently. Table 4 demonstrates that we are able to better capitalize on higher video and/or audio temporal resolution when using spatio-temporal augmentations compared to no spatio-temporal augmentation control models. For video frames, when varying the number of frames we observe higher performance for spatio-temporal augmentations compared to control **+7.71%**. Furthermore, we observe that by only increasing the number of

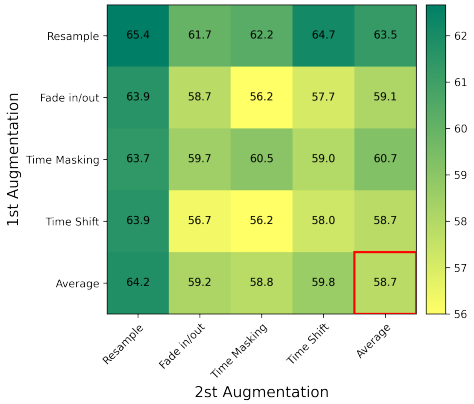

Figure 3: Top-1 accuracy of linear classifiers on representations pre-trained with contrastive loss on AVE dataset using one or pair of spatio-temporal augmentations for 500 epochs. The bottom-right element is a model trained with no spatio-temporal augmentation. The last row/column represents the average of each row/column, respectively. Diagonal of the matrix demonstrates the result when only one corresponding spatio-temporal augmentation applied during pre-training.

Table 3: Investigating the impact of location and alignment of the spatio-temporal augmentations. Top-1/5 accuracy of linear classifiers on representations pre-trained with contrastive loss using the two best performing spatio-temporal augmentations on AVE dataset for 500 epochs.

Augmentation Location	Augmentation			
	RE		RE + TS	
	Accuracy (%)		Accuracy (%)	
	top-1	top-5	top-1	top-5
No Aug.	58.70	90.55	58.70	90.55
Audio only	63.68	92.29	64.18	93.03
Video only	59.95	91.54	61.69	91.04
Both (+ alignment)	63.43	91.79	61.44	93.28
Both (+ no alignment)	**65.42**	**93.28**	64.18	91.29

frames from 4 to 32 during training, the gap in performance between control and spatio-temporal augmentation models increases by **+3.98%**. Alternatively, for audio temporal resolution we vary two parameters: spectrogram hop size and sampling rate (kHz). Although spectrogram hop size increases the amount of resolution of spectrograms across the temporal dimension, we find no benefit in increasing the spectrogram resolution. Because the gap in performance between control and spatio-temporal augmentation degrades when moving from 256 to 32 hop size (**-3.73%**). Similar to video frames, we observe that by increasing the audio sampling rate the gap in performance between control and spatio-temporal augmentation models increases when moving from 5.5 kHz to 44.1 kHz (**+2.48%**). In this paper, when not specified, the default number of frames is 8, spectrogram hop size is 128, and audio sampling rate is 44.1kHz.

4.3 Contrastive Learning Frameworks

Over the past few years, abundant number of contrastive frameworks has been proposed (He et al., 2020; Chen et al., 2020a; Grill et al., 2020; Chen & He, 2020). In this section, we explore different frameworks and investigate their performance on learning efficient audiovisual representations. Table 5 demonstrates the linear classifier top-1 accuracy of various contrastive frameworks on AVE dataset. We observe that frameworks (i.e. SimCLR and MoCoV3) that incorporate negative samples in their loss, generally require less training to reach competitive representations compared to frameworks (i.e. BYOL and SimSiam) that rely only on positive samples. Feichtenhofer et al. (2021b)

Table 4: Investigating the impact of temporal resolution of audio and video streams. Temporal resolution for video frames is adjusted through number of frames given to the model during pre-training on AVE dataset. For the audio signal, we adjusted the spectrogram hop size and sampling rate (kHz) during pre-training. Each section shows the effect of increasing temporal resolution with and without spatio-temporal augmentations on performance.

Temporal Augmentation	Temporal Resolution					
	lowest ⟶ *highest*					
Number of Frames	4	8	16	32	$\bar{x}$	
No Augmentation	57.71	58.71	58.46	54.73	57.40	**+7.71**
Resample	63.93	65.42	66.17	64.93	65.11	
$\Delta(x)$	6.22	6.71	7.71	10.2		
	+0.49		**+1.00**	**+2.49**		
Spectrogram Hop Size	256	128	64	32	$\bar{x}$	
No Augmentation	55.97	58.71	54.48	54.48	55.91	**+6.84**
Resample	64.18	65.42	62.44	58.96	62.75	
$\Delta(x)$	8.21	6.71	7.96	4.48		
	-1.50		**+1.25**	-3.48		
Audio Sampling Rate (kHz)	5.5	11	22	44.1	$\bar{x}$	
No Augmentation	53.48	54.98	60.20	58.71	56.84	**+3.92**
Resample	57.71	57.21	62.69	65.42	60.75	
$\Delta(x)$	4.23	2.23	2.49	6.71		
	-2.00		**+0.26**	**+4.22**		

showed similar trend with BYOL and MoCoV2 on K400 and UCF101 datasets trained only on video frames. Although SimCLR and MoCoV3 outperform other methods when pre-trained with ≤ 400 epochs, BYOL and SimSiam produce better representations when trained longer (i.e. 800 epochs). Lastly, SimSiam outperforms all methods with **71.64%** accuracy.

Table 5: Investigating the impact of number of self-supervised pre-training epochs and the type of self-supervised methods used. The same augmentations were used for all the models.

Pre-training Method	Top-1 Accuracy (%)			
	Pre-training Duration			
	100	200	400	800
Supervised	66.17	70.90	71.89	74.88
MoCoV3	58.21	59.95	62.44	66.92
SimCLR	59.45	63.18	66.17	68.66
BYOL	53.73	55.72	63.43	68.66
SimSiam	49.00	51.99	63.18	**71.64**

5 CONCLUSION

In this paper, with extensive and comprehensive experiments on various design choices and audio-visual augmentations, we proposed an effective new pipeline for audiovisual contrastive learning. Together, our results depict a promising path towards automated audiovisual integration to learn efficient audiovisual representations from video data.

REFERENCES

Haider Al-Tahan and Yalda Mohsenzadeh. Clar: Contrastive learning of auditory representations. In *International Conference on Artificial Intelligence and Statistics*, pp. 2530–2538. PMLR, 2021.

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks, 2020.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering, 2020a.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

Relja Arandjelović and Andrew Zisserman. Look, listen and learn, 2017.

Relja Arandjelović and Andrew Zisserman. Objects that sound, 2018.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video, 2016.

Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. 2020a.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. 2020b.

Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. 2017.

Alejandro Cartas, Jordi Luque, Petia Radeva, Carlos Segura, and Mariella Dimiccoli. Seeing and hearing egocentric actions: How much can we learn? 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. 2020.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020b.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. 2016.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. 2019.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3299–3309, June 2021a.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning, 2021b.

Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video, 2018.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. 2018.

Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. 2020.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. 2020.

Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning, 2019.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. 2020.

Dinesh Jayaraman and Kristen Grauman. Slow and steady feature analysis: higher order temporal coherence in video. 2016.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition, 2019.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition, 2021.

Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. 2017.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity, 2018.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features, 2018.

Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning, 2016.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. 2016.

Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. 2017.

Mandela Patrick, Yuki M. Asano, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations, 2020.

Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. 2020.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning, 2021.

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes, 2018.

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)*, pp. 247–263, 2018a.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos, 2018b.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. 2019.

Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. 2018.

Luyu Wang and Aaron van den Oord. Multi-format contrastive learning of audio representations. 2021.

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. 2015.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. 2018.

Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. 2020.

Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. 2019.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. 2016.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. 2017.

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions, 2019.