Distilling Audio-Visual Knowledge by Compositional Contrastive Learning (Supplementary)

Yanbei Chen[1], Yongqin Xian[2], A. Sophia Koepke[1], Ying Shan[3], Zeynep Akata[1,2,4]

[1]University of Tübingen [2]MPI for Informatics [3]Tencent PCG [4]MPI for Intelligent Systems

{yanbei.chen, a-sophia.koepke, zeynep.akata}@uni-tuebingen.de, yxian@mpi-inf.mpg.de

A. Additional Algorithmic Details

Algorithm A gives an overview of our compositional contrastive learning (CCL) algorithm for audio-visual distillation. From an information-theoretic point of view, CCL distills audio-visual knowledge from the teacher networks by maximising the mutual information between the student network $\theta_{\text{3D-CNN}}$ and the teacher networks $\theta_{\text{1D-CNN}}, \theta_{\text{2D-CNN}}$ and the composition functions $\mathcal{F}_{av}, \mathcal{F}_{iv}$. While the multi-class contrastive loss $\mathcal{L}_{nce}$ contrasts the feature similarity, the Jensen–Shannon divergence $\mathcal{L}_{JSD}$ contrasts the prediction similarity, which together maximise the similarities between the cross-modal positive pairs from the same class. Importantly, class labels are introduced into both loss terms and the composition functions (Figure A), thus ensuring to transfer the task-relevant knowledge to the student network.

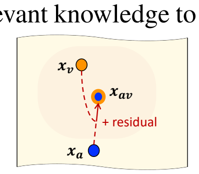

Figure A. Illustration of the compositional embedding x_{av}. The composition function uses residual learning to modify a teacher embedding x_a, which shifts x_a towards the video embedding x_v, resulting in the compositional embedding x_{av}. Given the classification constraint $\mathcal{L}_{ce}$, x_{av} is enforced to share to the same video class as x_v, thus closing the possible cross-modal semantic gap.

B. Additional Analysis and Results

Analysis of Cross-Modal Correspondence. Here, we first manually analyse the audio-video correspondence on UCF51. For each video, we compute the top-10 predicted audio classes using the audio network. For each video class, we compute the top-10 associated audio classes based on how frequently they are predicted as the top-10 audio classes. We summarise the video classes and their top-10 associated audio classes in Tables D, E, F, where we manually classify the audio-video correspondence as highly, weakly and not correlated, as summarised in Table A. Moreover, we give some qualitative examples of the image-video correspondence (Figure B). As shown, the visual cues in image frames are generally highly or weakly related to the video content, e.g. the *sea* image is weakly correlated with the video class *skijet* due to occlusion.

audio-video correspondence	# video classes	proportion (%)
highly	15	29.4
weakly	15	29.4
not	21	41.2

Table A. Statistics of the audio-video correspondence on UCF51.

Figure B. Image-video correspondence: videos (tagged in blue) from UCF. The red/green boxes mean that the visual cues in the image frames are highly/weakly correlated to the video classes.

Remark. Our analysis of the UCF51 dataset indicates the presence of a possible cross-modal semantic gap in multi-modal distillation. To empirically examine how CCL deal with this issue in practice, we evaluate CCL and its best competitor CRD using highly/weakly or not correlated audios for audio distillation. As Table B shows, on the 21 classes with not correlated audios, CRD performs on par with the baseline w/o distillation (+0.5% acc), while our CCL outperforms the baseline significantly (+3.1% acc). This suggests that our CCL can distill complementary information from audio even if it is uncorrelated with the video.

audio-video correspondence	baseline	CRD	**CCL**
weak/highly correlated (30 classes)	55.3	59.6	**65.5**
not correlated (21 classes)	61.0	61.5	**64.1**

Table B. Ablation study of audio distillation on UCF51.

Tabular Results on VGGSound. In Table C, we provide the tabular results of audio-visual distillation on the large-scale VGGSound dataset, which includes three contrastive learning methods (CRD, CMC, CCL) in comparison to the baseline for the video recognition and video retrieval tasks.

Task	Recognition		Retrieval			
Metric	Top1	Top5	R1	R5	R10	R20
baseline	19.1	20.3	22.1	38.5	47.0	55.8
CRD	19.8	41.6	22.0	39.2	47.8	56.3
CMC	12.6	30.0	18.3	34.7	43.1	52.1
CCL	**23.6**	**46.2**	**28.1**	**45.0**	**52.5**	**60.2**

Table C. Evaluating audio-visual distillation on VGGSound.

Algorithm A Compositional Contrastive Learning (Audio-Visual Distillation)

Require: Video dataset $\mathcal{D} = \{\mathbf{V}_i, y_i\}_{i=1}^{N}$, the corresponding image frames $\{\mathbf{I}_{ij}\}_{j=1}^{M_i}$ and audio recording $\mathbf{A}_i$ for each video.
Require: Trainable video student network $\theta_{\text{3D-CNN}}$. Pre-trained audio and image teacher networks $\theta_{\text{1D-CNN}}, \theta_{\text{2D-CNN}}$.

for $t = 1$ **to** *max_iter* **do**
  $x_a \leftarrow \theta_{\text{1D-CNN}}(\mathbf{A}_i), x_i \leftarrow \theta_{\text{2D-CNN}}(\mathbf{I}_{ij}), x_v \leftarrow \theta_{\text{3D-CNN}}(\mathbf{V}_i)$ {obtain unimodal audio, image, video embeddings}
  $x_{av} \leftarrow \mathcal{F}_{av}(x_a, x_v) x_{iv} \leftarrow \mathcal{F}_{iv}(x_i, x_v)$ {derive compositional embeddings}
  $\mathcal{L}_{ce}^{v} \leftarrow \mathcal{L}_{ce}(x_v, k), \mathcal{L}_{ce}^{av} \leftarrow \mathcal{L}_{ce}(x_{av}, k), \mathcal{L}_{ce}^{iv} \leftarrow \mathcal{L}_{ce}(x_{iv}, k)$ {compute video classification loss}
  $\mathcal{L}_{\text{audio}} \leftarrow \lambda \mathcal{L}_{nce}(x_v, x_a) + (1-\lambda)\mathcal{L}_{nce}(x_v, x_{av}) + JSD(P_v||P_{av})$ {compute audio distillation loss}
  $\mathcal{L}_{\text{image}} \leftarrow \lambda \mathcal{L}_{nce}(x_v, x_i) + (1-\lambda)\mathcal{L}_{nce}(x_v, x_{iv}) + JSD(P_v||P_{iv})$ {compute visual distillation loss}
  $\theta_{\text{3D-CNN}}^{t+1} \leftarrow \theta_{\text{3D-CNN}}^{t} - \eta \frac{\partial \mathcal{L}^v}{\partial \theta_{\text{3D-CNN}}}$, where $\mathcal{L}^v = \mathcal{L}_{ce}^{v} + \mathcal{L}_{\text{audio}} + \mathcal{L}_{\text{image}}$ {backprop on the video network}
  $\theta_{av}^{t+1} \leftarrow \theta_{av}^{t} - \eta \frac{\partial \mathcal{L}_{ce}^{av}}{\partial \theta_{av}}, \theta_{iv}^{t+1} \leftarrow \theta_{iv}^{t} - \eta \frac{\partial \mathcal{L}_{ce}^{iv}}{\partial \theta_{iv}}$ {backprop on the composition functions}
end for

Video Class	Top-10 Associated Audio Classes	Correlated
ApplyEyeMakeup	Speech; Inside, small room; Music; Female speech, woman speaking; Vehicle; Writing; Conversation; Narration, monologue; Animal; Rustle	not
ApplyLipstick	Speech; Music; Inside, small room; Vehicle; Animal; Female speech, woman speaking; Narration, monologue; Conversation; Musical instrument; Writing	not
Archery	Speech; Music; Vehicle; Arrow; Inside, small room; Outside, rural or natural; Animal; Car; Door; Bird	highly
BabyCrawling	Speech; Inside, small room; Music; Animal; Child speech, kid speaking; Babbling; Laughter; Domestic animals, pets; Vehicle; Crying, sobbing	weakly
BalanceBeam	Speech; Music; Vehicle; Outside, urban or manmade; Crowd; Inside, large room or hall; Inside, public space; Basketball bounce; Car; Animal	weakly
BandMarching	Music; Speech; Musical instrument; Drum; Percussion; Crowd; Orchestra; Brass instrument; Vehicle; Wood block	highly
BasketballDunk	Music; Speech; Vehicle; Basketball bounce; Outside, urban or manmade; Crowd; Car; Hip hop music; Slam; Singing	highly
BlowDryHair	Music; Speech; Vehicle; Inside, small room; Hair dryer; Vacuum cleaner; Car; Mechanical fan; Animal; Train	highly
BlowingCandles	Speech; Inside, small room; Music; Animal; Laughter; Chuckle, chortle; Snicker; Child speech, kid speaking; Inside, large room or hall; Domestic animals, pets	not
BodyWeightSquats	Speech; Music; Vehicle; Inside, small room; Male speech, man speaking; Narration, monologue; Animal; Musical instrument; Car; Conversation	not
Bowling	Speech; Music; Vehicle; Outside, urban or manmade; Train; Slam; Car; Animal; Inside, public space; Inside, large room or hall	weakly
BoxingPunchingBag	Speech; Music; Slam; Inside, large room or hall; Inside, small room; Thump, thud; Singing; Tap; Musical instrument; Vehicle	weakly
BoxingSpeedBag	Speech; Vehicle; Engine; Music; Car; Idling; Machine gun; Outside, urban or manmade; Engine starting; Motorcycle	weakly
BrushingTeeth	Speech; Inside, small room; Music; Animal; Toothbrush; Domestic animals, pets; Scratch; Rub; Vehicle; Child speech, kid speaking	highly
CliffDiving	Music; Speech; Vehicle; Musical instrument; Electronic music; Car; Rock music; Outside, urban or manmade; Guitar; Trance music	not
CricketBowling	Speech; Music; Vehicle; Outside, urban or manmade; Outside, rural or natural; Car; Animal; Basketball bounce; Slam; Inside, large room or hall	weakly
CricketShot	Speech; Music; Vehicle; Outside, urban or manmade; Arrow; Animal; Outside, rural or natural; Slam; Car; Thump, thud	weakly

Table D. Audio-video correspondence on the UCF51 classes. Video classes (1˜17), the top-10 associated audio classes, and the audio-video correlation: highly, weakly, or not correlated. Note: audio events highly/weakly correlated with the video are highlighted in red/green.

Video Class	**Top-10 Associated Audio Classes**	**Correlated**
CuttingInKitchen	Music; Speech; Inside, small room; Chopping (food); Dishes, pots, and pans; Wood; Animal; Chop; Vehicle; Cutlery, silverware	highly
FieldHockeyPenalty	Speech; Vehicle; Music; Outside, urban or manmade; Basketball bounce; Car; Animal; Crowd; Outside, rural or natural; Hubbub, speech noise, speech babble	weakly
FloorGymnastics	Music; Speech; Crowd; Cheering; Vehicle; Inside, large room or hall; Children shouting; Outside, urban or manmade; Singing; Whoop	not
FrisbeeCatch	Speech; Music; Vehicle; Outside, urban or manmade; Singing; Car; Pop music; Hubbub, speech noise, speech babble; Boat, Water vehicle; Crowd	not
FrontCrawl	Speech; Vehicle; Music; Water; Stream; Car; Boat, Water vehicle; Outside, urban or manmade; Splash, splatter; Outside, rural or natural	weakly
Haircut	Speech; Music; Inside, small room; Vehicle; Musical instrument; Animal; Inside, large room or hall; Electronic music; Outside, urban or manmade; Female speech, woman speaking	not
HammerThrow	Music; Speech; Vehicle; Outside, urban or manmade; Car; Male speech, man speaking; Animal; Musical instrument; Outside, rural or natural; Basketball bounce	not
Hammering	Music; Speech; Hammer; Whack, thwack; Chop; Tools; Inside, small room; Musical instrument; Vehicle; Glass	highly
HandstandPushups	Speech; Music; Vehicle; Animal; Inside, small room; Musical instrument; Singing; Domestic animals, pets; Car; Pink noise	not
HandstandWalking	Speech; Music; Vehicle; Animal; Outside, urban or manmade; Car; Inside, small room; Musical instrument; Singing; Domestic animals, pets	not
HeadMassage	Speech; Music; Vehicle; Outside, urban or manmade; Inside, small room; Musical instrument; Singing; Animal; Inside, large room or hall; Car	not
IceDancing	Music; Speech; Musical instrument; Vehicle; Orchestra; Singing; Theme music; Outside, urban or manmade; Television; Guitar	not
Knitting	Speech; Inside, small room; Music; Writing; Animal; Vehicle; Female speech, woman speaking; Conversation; Air conditioning; Narration, monologue	not
LongJump	Speech; Music; Outside, urban or manmade; Vehicle; Car; Basketball bounce; Inside, public space; Crowd; Hubbub, speech noise, speech babble; Run	weakly
MoppingFloor	Speech; Inside, small room; Music; Animal; Vehicle; Inside, large room or hall; Male speech, man speaking; Television; Narration, monologue; Domestic animals, pets	not
ParallelBars	Speech; Music; Crowd; Inside, large room or hall; Inside, public space; Outside, urban or manmade; Basketball bounce; Cheering; Slam; Vehicle	weakly
PlayingCello	Music; Musical instrument; Bowed string instrument; Cello; String section; Violin, fiddle; Double bass; Classical music; Orchestra; Piano	highly

Table E. Audio-video correspondence on the UCF51 classes. Video classes (18˜34), the top-10 associated audio classes, and the audio-video correlation: highly, weakly, or not correlated. Note: audio events highly/weakly correlated with the video are highlighted in red/green.

Video Class	**Top-10 Associated Audio Classes**	**Correlated**
PlayingDaf	Music; Drum; Musical instrument; Percussion; Drum kit; Bass drum; Snare drum; Drum roll; Wood block; Rimshot	highly
PlayingDhol	Music; Drum; Musical instrument; Percussion; Drum kit; Bass drum; Wood block; Speech; Snare drum; Tabla	highly
PlayingFlute	Musical instrument; Music; Flute; Wind instrument, woodwind instrument; Classical music; Inside, small room; Bowed string instrument; Piano; Violin, fiddle; Speech	highly
PlayingSitar	Music; Musical instrument; Sitar; Plucked string instrument; Classical music; Carnatic music; Bowed string instrument; Speech; Tabla; Music of Asia	highly
Rafting	Music; Speech; Vehicle; Singing; Musical instrument; Waves, surf; Ocean; Car; Waterfall; Guitar	weakly
ShavingBeard	Speech; Inside, small room; Music; Vehicle; Animal; Inside, large room or hall; Electric shaver, electric razor; Outside, urban or manmade; Electric toothbrush; Buzz	highly
Shotput	Speech; Music; Vehicle; Outside, urban or manmade; Animal; Car; Outside, rural or natural; Musical instrument; Hubbub, speech noise, speech babble; Singing	not
SkyDiving	Music; Musical instrument; Singing; Punk rock; Rock music; Heavy metal; Grunge; Progressive rock; Angry music; Rock and roll	not
SoccerPenalty	Speech; Outside, urban or manmade; Music; Vehicle; Basketball bounce; Crowd; Slam; Male speech, man speaking; Car; Inside, public space	weakly
StillRings	Speech; Music; Vehicle; Outside, urban or manmade; Car; Inside, public space; Basketball bounce; Crowd; Inside, large room or hall; Slam	not
SumoWrestling	Speech; Music; Crowd; Inside, large room or hall; Inside, public space; Outside, urban or manmade; Cheering; Chatter; Basketball bounce; Slam	weakly
Surfing	Music; Musical instrument; Vehicle; Speech; Rock music; Rock and roll; Punk rock; Guitar; Singing; Car	not
TableTennisShot	Speech; Music; Ping; Animal; Tap; Inside, small room; Inside, large room or hall; Vehicle; Whack, thwack; Bouncing	highly
Typing	Typing; Speech; Computer keyboard; Typewriter; Vehicle; Inside, small room; Music; Animal; Engine; Sewing machine	highly
UnevenBars	Music; Speech; Outside, urban or manmade; Vehicle; Crowd; Basketball bounce; Car; Cheering; Slam; Children shouting	not
WallPushups	Speech; Music; Inside, small room; Animal; Vehicle; Narration, monologue; Male speech, man speaking; Conversation; Female speech, woman speaking; Musical instrument	not
WritingOnBoard	Speech; Inside, small room; Music; Male speech, man speaking; Narration, monologue; Inside, large room or hall; Chopping (food); Writing; Chop; Conversation	weakly

Table F. Audio-video correspondence on the UCF51 classes. Video classes (35˜51), the top-10 associated audio classes, and the audio-video correlation: highly, weakly, or not correlated. Note: audio events highly/weakly correlated with the video are highlighted in red/green.