

Learning From Music to Visual Storytelling of Shots: A Deep Interactive Learning Mechanism

Jen-Chun Lin¹, Wen-Li Wei¹, Yen-Yu Lin², Tyng-Luh Liu¹, Hong-Yuan Mark Liao^{1,3}

¹Institute of Information Science, Academia Sinica, Taiwan,

²Department of Computer Science, National Chiao Tung University, Taiwan,

³Department of Computer Science and Information Engineering, Providence University

{jenchunlin,lilijinjin}@gmail.com;lin@cs.nctu.edu.tw;liutyng@iis.sinica.edu.tw;liao@iis.sinica.edu.tw

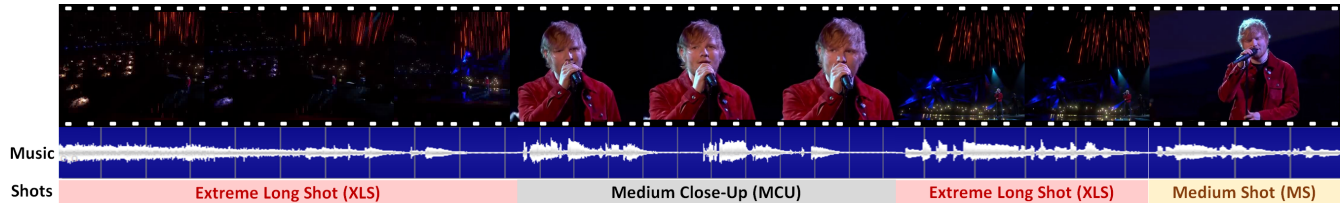


Figure 1: An example of visual storytelling for the song “Supermarket Flowers” by Ed Sheeran live at the BRIT Awards 2018. The director sequentially uses the extreme long shot, medium close-up, extreme long shot, and medium shot to expand the storytelling potential in the beginning of the song. The video demo can be found at <https://sites.google.com/site/m2vsdil/>.

ABSTRACT

Learning from music to visual storytelling of shots is an interesting and emerging task. It produces a coherent visual story in the form of a shot type sequence, which not only expands the storytelling potential for a song but also facilitates automatic concert video mashup process and storyboard generation. In this study, we present a deep interactive learning (DIL) mechanism for building a compact yet accurate sequence-to-sequence model to accomplish the task. Different from the one-way transfer between a pre-trained teacher network (or ensemble network) and a student network in knowledge distillation (KD), the proposed method enables collaborative learning between an ensemble teacher network and a student network. Namely, the student network also teaches. Specifically, our method first learns a teacher network that is composed of several assistant networks to generate a shot type sequence and produce the soft target (shot types) distribution accordingly through KD. It then constructs the student network that learns from both the ground truth label (hard target) and the soft target distribution to alleviate the difficulty of optimization and improve generalization capability. As the student network gradually advances, it turns to feed back knowledge to the assistant networks, thereby improving the teacher network in each iteration. Owing to such interactive designs, the DIL mechanism bridges the gap between the teacher and student networks and produces more superior capability for both networks. Objective and subjective experimental results demonstrate that both the teacher and student networks can generate

more attractive shot sequences from music, thereby enhancing the viewing and listening experience.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence.**

KEYWORDS

visual storytelling, ensemble network, knowledge distillation

ACM Reference Format:

Jen-Chun Lin¹, Wen-Li Wei¹, Yen-Yu Lin², Tyng-Luh Liu¹, Hong-Yuan Mark Liao^{1,3}. 2020. Learning From Music to Visual Storytelling of Shots: A Deep Interactive Learning Mechanism. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, October 12–16, 2020, Seattle, WA, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3394171.3413985>

1 INTRODUCTION

Recently, a challenging research topic has been proposed, namely learning from music to visual storytelling of shots [26]. It aims to interpret music with an appropriate and near-professional shot type (defined in the language of film, see Table 1) sequence for visual storytelling in concert videos. Figure 1 illustrates how visual storytelling is achieved in making an official concert video. In particular, it shows that the director sequentially uses the extreme long shot, medium close-up, extreme long shot, and medium shot in the beginning of the song to express the emotion and expand the storytelling potential of the song. Therefore, understanding how to properly employ shots is crucial in carrying out an automated concert video mashup process and storyboard generation [1, 10, 15, 21, 25–27].

The task of music to visual storytelling of shots translation is challenging due to the difficulty of modeling the large musical variance in an ordered song structure and preserving the long-term coherence among multiple shots. Existing recurrent neural network (RNN)-based approaches only consider the adjacent memories in a sequence [13, 17], which may not be able to effectively learn the long-term dependencies over large and varied temporal

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.







MM '20, October 12–16, 2020, Seattle, WA, USA

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-7988-5/20/10...\$15.00

<https://doi.org/10.1145/3394171.3413985>

Table 1: The definition of six types of shots [1, 15].

					
Close-Up (CU)	Medium Close-Up (MCU)	Medium Shot (MS)	Medium Long Shot (MLS)	Long Shot (LS)	Extreme Long Shot (XLS)
A Close-Up is used to show emotion on the subject's face. That is, the face occupies most of the screen (image).	A Medium Close-Up contains a subject's head and shoulders completely.	A Medium Shot contains a subject from the waist to the top of the head.	A Medium Long Shot would contain a subject from his/her knees to the top of the head.	A Long Shot would contain a subject's entire body from the top of the head to the bottom of the feet.	An Extreme Long Shot covers a large area or landscape. It would be hard to see any reactions /emotion from people in the shot since they are too far away.

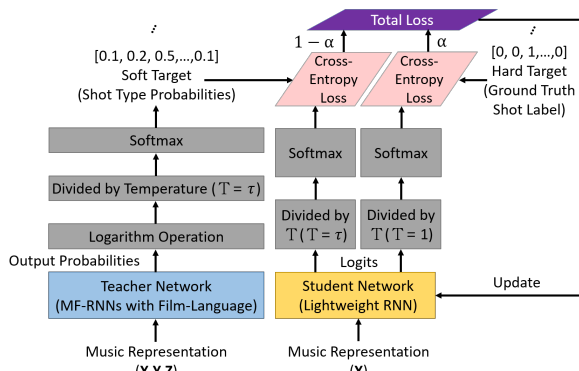


Figure 2: Illustration of the proposed distillation-based teacher-to-student learning framework.

intervals in a concert video. To address this difficulty, a probabilistic-based ensemble model has recently been proposed, termed as multi-resolution fused recurrent neural networks (MF-RNNs) with film-language [26]. This model not only explores the varied temporal resolution RNNs, but also relates the statistical dependencies among them, and further integrates a film-language model to improve music-to-shot translation performance. However, despite the effectiveness in translation, such an ensemble model has a high computational complexity, which limits the applicability in practice. In addition, each temporal resolution RNN used in MF-RNNs with film-language is learned directly from the ground truth (shot) label (so-called hard target), which can only provide very limited information.

To address the aforementioned issues, we propose a novel collaborative model distillation mechanism called deep interactive learning (DIL) for building an effective lightweight RNN. Different from the one-way transfer between a pre-trained teacher network (or ensemble network) and a student network in knowledge distillation (KD) [8], the proposed DIL mechanism enables collaborative learning between an ensemble teacher network and a student network. Namely, the student network also teaches. Thus, the DIL mechanism can be regarded as a bi-directional KD approach, which enables iterative interaction between the teacher and student networks during KD. The proposed mechanism is developed based on the observation of KD experiments in MF-RNNs with film-language. That is, we use a lightweight RNN as a student network to mimic the output behavior of MF-RNNs with film-language (teacher network), see Figure 2. The architecture of MF-RNNs with film-language is given in Figure 3. We empirically find that the performance of the student network is better than the three temporal resolution RNNs (called assistant networks) used in MF-RNNs with film-language. This observation inspires us that the knowledge of the student

network can be fed back to each assistant network, thus upgrading the teacher network.

In the proposed DIL mechanism, learning comes from two aspects: teacher-to-student and student-to-teacher. Regarding teacher-to-student learning, as shown in Figure 2, the process starts with a powerful pre-trained teacher network, MF-RNNs with film-language, and then performs knowledge transfer to a student network (lightweight RNN) through KD. The teacher network mainly consists of three assistant networks that are initially trained by only using the ground truth labels (hard target). The student network is trained by further considering the soft target distribution (distilled from teacher network) that aligns probabilities of output classes to the pre-trained teacher network, as shown in Figure 2. In student-to-teacher learning, as shown in Figure 4, we turn to distill the knowledge (soft target distribution) from the student network and then transfer the knowledge to each of assistant networks (temporal resolution RNNs), thereby upgrading the teacher network. Learning in this way, *i.e.*, finding and matching the other most likely classes (shot types) for each training instance according to their peers, can increase the posterior entropy of the learned network [4, 18], which helps both teacher and student networks converge to more robust minima and achieve better generalization to testing data. Thus, repeating such a DIL mechanism gradually improves the performance of both student and teacher networks.

We summarize our main contributions as follows:

- This study proposes a novel DIL mechanism, which provides a simple yet effective way to improve the generalization capability of a network through collaborative learning with the other network.
- Extensive experimental results show that compared with the MF-RNNs with film-language [26] and the conventional KD technique [8], the proposed DIL mechanism outperforms the competing methods, and achieves the state-of-the-art results by using a much smaller network.

2 RELATED WORK

Recently, learning compact yet accurate models has become active and has been approached in a variety of ways including model compression [5], pruning [9], binarisation [19] and model distillation [8]. This work targets at collaborative model distillation. In the following, we review the studies on KD and collaborative learning.

Distillation-based model compression is developed based on the observation that small networks often have the same representation capabilities as large networks [2, 3]. However, compared to large networks, small networks are less effective in training and seeking the optimal parameters to achieve the desired function. Such a problem results from the difficulty of optimization, rather than the size of the network [2]. KD has recently been proposed to

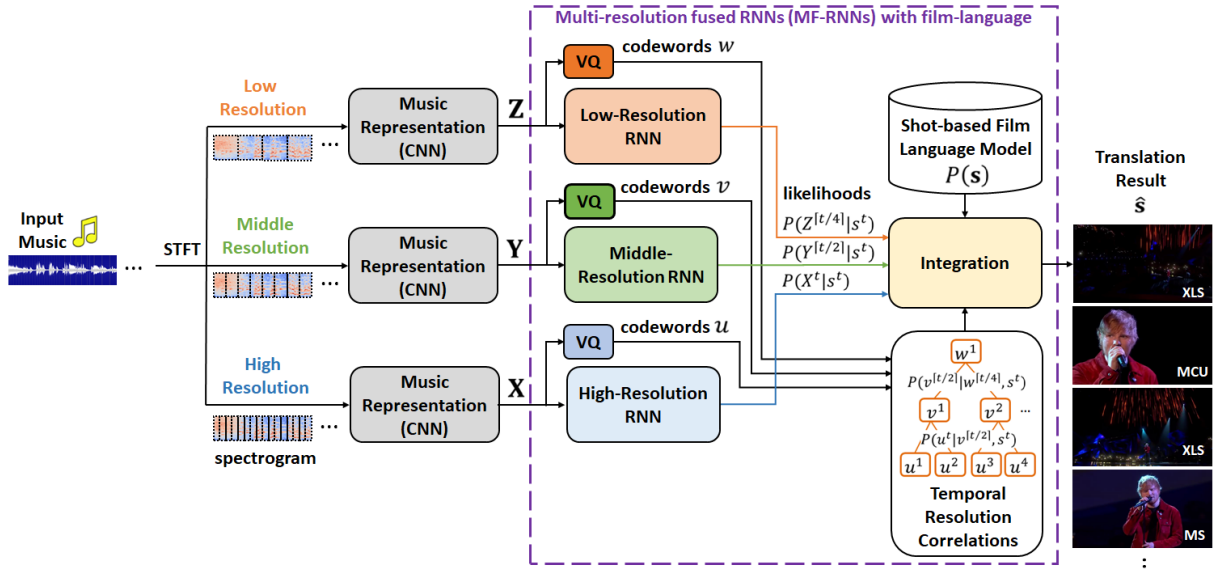


Figure 3: Illustration of the adopted teacher network, MF-RNNs with the film-language framework, for music to visual storytelling of shots translation [26]. Here STFT stands for short-time Fourier transform, and VQ stands for vector quantization.

make a small network easy to train. It typically starts with a powerful (deep and/or wide) teacher (or ensemble) network, and then trains a smaller (relatively shallow and/or narrow) student network to mimic the behavior (e.g., output class probabilities (so-called soft target) and/or feature representations) of the teacher network [2, 8, 20]. It turns out that learning to mimic the behavior of a teacher network is easier to optimize than learning directly from ground truth labels. Some intuition about why it works is due to the additional supervision and regularization of the higher entropy soft targets [3, 8, 12]. Ba and Caruana [2] train a shallow student network to mimic a deep teacher network by matching logits via the L_2 loss. Hinton et al. [8] generalize the KD technique by training a student network to predict soft target distribution provided by a teacher network. Sau and Balasubramanian [22] propose the logit perturbation mechanism, which adds random perturbations to logit outputs for simulating the process of learning from multiple teachers. Unlike the past, Yuan et al. [30] first train the student network in the normal way to obtain a pre-trained model, and then use it as a teacher to train itself by transferring soft targets. Instead of using soft targets, some studies use the hidden layer information of a teacher network to train the student network. For example, Romero et al. [20] propose to use intermediate representations learned by a teacher model as hints to improve the training process and final performance of the student. Yim et al. [28] define the distilled knowledge to be transferred in terms of flow between layers, which is calculated by computing the inner product between feature maps from two layers. Liu et al. [14] further propose an end-to-end KD framework that combines feature-level distillation and class-level distillation for multi-label image classification. However, despite the effectiveness of KD, such one-way knowledge transfer prevents a teacher network from improving with the student network. As English proverb mentions “*He who teaches, learns.*” Feedback from students also helps a teacher upgrade the knowledge. Once the teacher network upgrades, its knowledge in turn can further enrich

the student network. Thus, the student and teacher networks can grow together in such interaction. In this work, we develop the DIL mechanism that enables teacher and student networks to learn and grow from each other.

Related studies on collaborative learning include dual learning [6] and deep mutual learning [31]. Dual learning allows two cross-lingual translation models teach each other interactively, but it is applicable to special translation problems where an unconditional within-language model is available for assessing the quality of predictions, and ultimately provides the supervision for learning. In dual learning, different models address diverse learning tasks. In our DIL mechanism, the tasks are identical. In deep mutual learning, a learning scenario begins with a pool of untrained student networks that are simultaneously derived to solve the task together. Such a learning scenario differs from ours in that our DIL mechanism starts with a pre-trained ensemble teacher network and then learns interactively with a student network to make both sides grow.

3 DEEP INTERACTIVE LEARNING

In this section, we elaborate the proposed deep interactive learning (DIL) mechanism for translating music to visual storytelling of shots. We first revisit the adopted teacher network, MF-RNNs with film-language [26]. Then, we will introduce the two learning processes of DIL in detail, including teacher-to-student and student-to-teacher knowledge distillation.

3.1 Revisiting MF-RNNs with Film-Language

Given the music representation sequences of high, middle, and low temporal resolutions X , Y , and Z , the goal is to decide the shot sequence \hat{s} corresponding to the high temporal resolution, which can be cast as follows:

$$\hat{s} = \arg \max_{s \in S} P(s | X, Y, Z) \propto P(X, Y, Z | s)P(s), \quad (1)$$

where S denotes the set of all possible high temporal resolution shot sequences. $P(s)$ is the *a priori* probability of film-language,

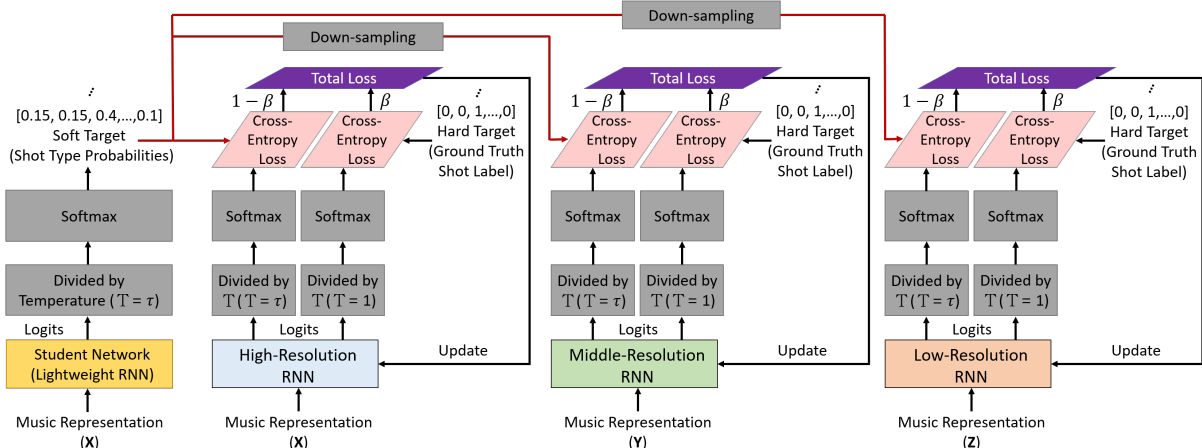


Figure 4: Illustration of the proposed distillation-based student-to-teacher learning framework.

which is calculated through statistical shot transitions from official concert videos in the training set. For estimating $P(s)$, a bigram language model is used via

$$P(s) = P(s^1) \prod_{t=2}^T P(s^t | s^{t-1}), \quad (2)$$

where T is the number of frames in the high temporal resolution sequence. In the following, we consider the estimation of the joint probability $P(X, Y, Z | s)$ to infer the optimal shot sequence \hat{s} in (1).

In practice, estimating $P(X, Y, Z | s)$ is difficult or even infeasible. To tackle the issue, Wei et al. [26] adopt the principle of maximum entropy proposed by Pan et al. [16], and define the joint probability as

$$P(X, Y, Z | s) \stackrel{def}{=} \prod_{t=1}^T P(X^t | s^t) P(Y^{\lceil \frac{t}{2} \rceil} | s^t) P(Z^{\lceil \frac{t}{4} \rceil} | s^t) \times \frac{P(u^t | v^{\lceil \frac{t}{2} \rceil}, s^t) P(v^{\lceil \frac{t}{2} \rceil} | w^{\lceil \frac{t}{4} \rceil}, s^t)}{P(u^t | s^t) P(v^{\lceil \frac{t}{2} \rceil} | s^t)}, \quad (3)$$

where $P(X^t | s^t)$, $P(Y^{\lceil \frac{t}{2} \rceil} | s^t)$, and $P(Z^{\lceil \frac{t}{4} \rceil} | s^t)$ are the likelihood of music representations of high, middle, and low temporal resolutions, respectively. Note that X , Y , and Z are the representations of 188 music tags encoded by a convolutional neural network, *i.e.*, those corresponding to 1-sec, 2-sec, and 4-sec per music frame, respectively [11, 26]. The symbol $\lceil \cdot \rceil$ represents the ceiling function. It is used to locate a frame of a middle or low temporal resolution that corresponds to the high temporal resolution. Take $t=1$ as an example. The first frame of the middle ($\lceil \frac{1}{2} \rceil$) and that of the low ($\lceil \frac{1}{4} \rceil$) temporal resolutions are located to help translate the shot type of the first frame of the high temporal resolution.

The remaining in the right hand side of (3) is the correlation term of multiple temporal resolutions. Because X , Y , and Z are continuous, it is infeasible to collect sufficient training data and use these data to figure out the statistical dependencies among the three, under the joint condition s . Therefore, k -means clustering is adopted to construct a codebook with which vector quantization is used to derive the corresponding codewords for X , Y , and Z , which are denoted by u , v , and w respectively in (3). $P(u^t | v^{\lceil \frac{t}{2} \rceil}, s^t)$, $P(v^{\lceil \frac{t}{2} \rceil} | w^{\lceil \frac{t}{4} \rceil}, s^t)$, $P(u^t | s^t)$, and $P(v^{\lceil \frac{t}{2} \rceil} | s^t)$ in (3) are calculated by statistical co-occurrence dependencies over all the training data:

$$P(u^t | v^{\lceil \frac{t}{2} \rceil}, s^t) = \frac{\text{Count}(u^t, v^{\lceil \frac{t}{2} \rceil}, s^t)}{\text{Count}(v^{\lceil \frac{t}{2} \rceil}, s^t)}, \quad (4)$$

$$P(v^{\lceil \frac{t}{2} \rceil} | w^{\lceil \frac{t}{4} \rceil}, s^t) = \frac{\text{Count}(v^{\lceil \frac{t}{2} \rceil}, w^{\lceil \frac{t}{4} \rceil}, s^t)}{\text{Count}(w^{\lceil \frac{t}{4} \rceil}, s^t)}, \quad (5)$$

$$P(u^t | s^t) = \frac{\text{Count}(u^t, s^t)}{\text{Count}(s^t)}, \quad (6)$$

and

$$P(v^{\lceil \frac{t}{2} \rceil} | s^t) = \frac{\text{Count}(v^{\lceil \frac{t}{2} \rceil}, s^t)}{\text{Count}(s^t)}. \quad (7)$$

The probabilities of co-occurrence dependencies can be regarded as the *a priori* knowledge to help translate music in the test phase. We use a dedicated RNN to separately model each of the three temporal resolutions and estimate their likelihood during implementation. Note that the output by an RNN is a posterior probability. Hence, we approximate the first likelihood in (3) by $P(s^t | X^t)/P(s^t)$, the second one by $P(s^t | Y^{\lceil \frac{t}{2} \rceil})/P(s^t)$, and the third one by $P(s^t | Z^{\lceil \frac{t}{4} \rceil})/P(s^t)$. Equations (2)-(7) specify how $P(s | X, Y, Z)$ in (1) is estimated by MF-RNNs with film-language in the test phase, which is also illustrated in Figure 3. It follows that the shot sequence s with the maximal posterior probability is selected as the translation result \hat{s} .

3.2 Proposed DIL Mechanism

Although MF-RNNs with film-language can deal with large musical variance in an ordered song structure and preserve long-term coherence among multiple shots by integrating multi-resolution fused RNNs and the film-language model, deploying it in practical applications is not easy. We remark that MF-RNNs with film-language is computationally heavy when performing music-to-shot translation, since it requires reading music information for at least 4 seconds (low temporal resolution) and integrating multiple deep-net models for 1 second (high temporal resolution) translation. In addition, each temporal resolution RNN used in MF-RNNs with film-language is learned directly from ground truth (shot) label that can provide very limited information. To make the network efficient and effective, we propose a DIL mechanism for building a lightweight RNN (student network) through collaborative learning with MF-RNNs with film-language (teacher network). The DIL mechanism

involves two learning processes, namely teacher-to-student, and student-to-teacher.

teacher-to-student The teacher-to-student learning process mainly involves two operations: pre-training an ensemble teacher network, and distilling the knowledge in teacher network to the student network, as shown in Figure 2. Motivated by the superior performance of MF-RNNs with film-language [26], we construct the teacher network based on the architecture of MF-RNNs with film-language. Specifically, we use three dedicated RNNs, *i.e.*, hRNN, mRNN, and lRNN, each of which contains 3 hidden layers with 256, 128, and 64 LSTM units to model music representations of high, middle, and low temporal resolutions, respectively. Each temporal resolution RNN (assistant network) is trained by using back-propagation through time, namely the BPTT algorithm with the objective of softmax cross entropy under the supervision of the ground truth (shot type) label. Then, we statistically estimate the co-occurrence dependencies and film-language to collaborate with the three temporal resolution RNNs to form the MF-RNNs with film-language for music-to-shot translation.

For knowledge distillation, we first build a lightweight RNN as our student network. Similar to the architecture of the above-mentioned temporal resolution RNNs, the student network is composed of 3 hidden layers, each with 256, 128, and 64 LSTM units, respectively. Thus, the model complexity of the student network is reduced to less than one-third of the MF-RNNs with film-language. For distillation, the student network takes the music representation of the high temporal resolution as input and mimic the output behavior of the teacher network that reads music representations from three temporal resolutions. Specifically, knowledge is transferred from the teacher network to the student network by minimizing the loss function which takes into consideration of the distribution of class (shot type) probabilities predicted by the teacher network, as shown in Figure 2. However, such a distribution typically peaks very sharply at the correct class, while the probability in other classes is close to zero. Therefore, it does not provide much information beyond the ground truth label. To address this issue, we use the softmax temperature [8] to control the variance of the probabilities between different classes and generate soft target distribution $p^{teacher}$ with

$$p_i^{teacher} = \frac{\exp(z_i^{teacher}/\mathbb{T})}{\sum_j \exp(z_j^{teacher}/\mathbb{T})}, \quad (8)$$

where $p_i^{teacher}$ represents the probability of the softmax temperature of the i -th class in the teacher network. Since our teacher network needs to integrate multiple probabilistic models such as temporal resolution RNNs and film-language model to produce the translation result, unlike [8], we apply the operation of softmax temperature directly to the output probabilities of the teacher network, instead of the logits of each integrated model. To this end, $z_i^{teacher}$ is defined as $z_i^{teacher} = \log(o_i^{teacher})$ where $o_i^{teacher}$ is the output probability of the teacher network of the i -th class. \mathbb{T} represents the temperature parameter. As \mathbb{T} grows, the probability distribution generated by the softmax function becomes softer, providing more information about classes that the teacher network finds more relevant to the predicted class.

It has been proven that combining the loss of the soft target distribution and ground truth label (so-called hard target) [8, 23] is

beneficial to student network learning. Thus, the total loss function for student network learning is given by

$$\mathcal{L}(\mathbf{X}; \theta) = \alpha \times \mathcal{H}(g^h, \sigma(z^{student}; \mathbb{T} = 1)) + (1 - \alpha) \times \mathcal{H}(\sigma(z^{teacher}; \mathbb{T} = \tau), \sigma(z^{student}; \mathbb{T} = \tau)), \quad (9)$$

where \mathbf{X} is the input music representation sequence of high temporal resolution, θ is the parameter set of the student network, α set to 0.5 is a non-negative hyper-parameter, \mathcal{H} is the cross-entropy loss function, g^h is the ground truth label of high temporal resolution, σ is the softmax function parameterized by the temperature \mathbb{T} , $z^{student}$ is the logits of the student network, and constant τ is set to 20 in all experiments. The detailed learning procedure is given in Figure 2. After learning, the student network can be directly used to get the shot sequence from the input music representation sequence of the high temporal resolution, regardless of the middle and low temporal resolutions, hence greatly improving efficiency.

student-to-teacher For student-to-teacher learning, we aim to upgrade the teacher’s knowledge by distilling the knowledge from the student network to each of the assistant networks integrated in the teacher network. Similar to teacher-to-student learning, we use the softmax temperature to generate a soft target distribution $p^{student}$ from the logits of the student network according to

$$p_i^{student} = \frac{\exp(z_i^{student}/\mathbb{T})}{\sum_j \exp(z_j^{student}/\mathbb{T})}, \quad (10)$$

where $p_i^{student}$ represents the probability of the softmax temperature of the i -th class in the student network. $z_i^{student}$ is the logit of the student network of the i -th class, and \mathbb{T} is the temperature. Since the soft target distribution generated by the student network represents a translation result of the music representation of the high temporal resolution, for two assistant networks *i.e.*, the middle- and low-resolution RNNs learning, as shown in Figure 4, we aggregate the soft target distribution of frames of the high temporal resolution in the student network through down-sampling. To this end, the probability of the softmax temperature of the i -th class in the m -th middle temporal resolution frame is evaluated via

$$p_i^{mid}(m) = \frac{1}{2} \sum_{t=2m-1}^{2m} p_i^{student}(t), \quad (11)$$

where $p_i^{student}(t)$ represents the probability of the softmax temperature of the i -th class in the t -th high temporal resolution frame resulted from the student network. Similarly, for low temporal resolution, the probability of the softmax temperature of the i -th class in the l -th low temporal resolution frame can be represented as

$$p_i^{low}(l) = \frac{1}{4} \sum_{t=4l-3}^{4l} p_i^{student}(t). \quad (12)$$

The regularization factors (*i.e.*, 1/2 and 1/4) are set according to a multiple of three temporal resolutions (*i.e.*, 4-sec, 2-sec, and 1-sec per frame). For middle and low temporal resolutions, the soft target distributions p^{mid} and p^{low} can be obtained by p_i^{mid} and p_i^{low} over all classes, respectively. For assistant network learning, the total loss function of the high-/middle-/low-resolution RNN is defined as

$$\mathcal{L}(\mathbf{I}; \pi) = \beta \times \mathcal{H}(g, \sigma(z^{assistant}; \mathbb{T} = 1)) + (1 - \beta) \times \mathcal{H}(\sigma(z^{student}; \mathbb{T} = \tau), \sigma(z^{assistant}; \mathbb{T} = \tau)), \quad (13)$$

Algorithm 1: Deep Interactive Learning (DIL)

Input : Music representation sequences $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$,
Ground truth labels g^h, g^m, g^l .

Initialize: $k = 0$, indexing the k -th DIL iteration,
assistant network parameters $\pi_k^h, \pi_k^m, \pi_k^l$,
student network parameters θ_k .

Output : Student network parameters θ .

- 1 Train three assistant networks $\pi_k^h, \pi_k^m, \pi_k^l$ with $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$,
 g^h, g^m , and g^l .
- 2 Compute the teacher network by Eq (1).
- 3 Distill knowledge $p^{teacher}$ from the teacher network by Eq
(8).
- 4 Train the student network θ_k via \mathbf{X}, g^h , and $p^{teacher}$.
- 5 $\theta \leftarrow \theta_k$.
- 6 **repeat**
- 7 $k = k + 1$.
- 8 Distill knowledge $p^{student}$ from the student network by
Eq (10).
- 9 Compute p^{mid} and p^{low} by Eq (11) and Eq (12).
- 10 Update $\pi_k^h, \pi_k^m, \pi_k^l$ with $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, g^h, g^m, g^l$,
 $p^{student}$, p^{mid} , and p^{low} , respectively.
- 11 Re-compute the teacher network by Eq (1).
- 12 Distill knowledge $p^{teacher}$ from the teacher network by
Eq (8).
- 13 Update the student network θ_k via \mathbf{X}, g^h , and $p^{teacher}$.
- 14 **if** the validation error of θ_k is less than θ **then**
- 15 | $\theta \leftarrow \theta_k$.

until the validation error of θ_k is not less than θ

where \mathbf{I} is the input music representation sequence, which can be replaced by \mathbf{X}, \mathbf{Y} or \mathbf{Z} according to high, middle or low temporal resolution, respectively. π is the assistant network parameters, which can be replaced by π^h, π^m or π^l for high-, middle-, or low-resolution RNN, respectively. β is a non-negative hyper-parameter, and is set to 0.5. \mathcal{H} is the cross-entropy loss function. g is the ground truth label, and it can be replaced by g^h, g^m or g^l for high, middle or low temporal resolution, respectively. σ is the softmax function parameterized by the temperature \mathbb{T} . $z^{assistant}$ and $z^{student}$ are the logits of the assistant and student networks, respectively, and τ is also set to 20 in all experiments. Note that for middle- and low-resolution RNNs learning, the output (soft target distribution) of the term $\sigma(z^{student}; \mathbb{T} = \tau)$ needs to be further aggregated using formulas (11) and (12).

The detailed learning procedure is given in Figure 4. Under the supervision of the soft target distribution of the student network feedback, the generalization capability of each assistant network can be further improved. Once the assistant networks are upgraded, the performance of the teacher network will be enhanced accordingly. When the teacher network is upgraded, its knowledge will be transferred to the student network in the next iteration through the teacher-to-student learning process. These two learning processes are repeated until no further supportable evidence can be detected. All in all, this collaborative mechanism allows DIL to

drive the two coupled teacher and student networks to learn and grow from each other. Algorithm 1 summarizes the procedure of the DIL mechanism.

4 EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed DIL mechanism, we conduct experiments on a set of official concert videos downloaded from YouTube links provided by [26]. In total, 60 official concert videos, each of which belongs to a complete song, are collected from 51 live concerts. Among them, 45 official concert videos are used for training, including 12,300 (high-resolution), 6,150 (middle-resolution), and 3,075 (low-resolution) labeled music frames [26]. Note that eight shot types (labels) are considered in this database, including CU, MCU, MS, MLS, LS, XLS (see Table 1), and two additional variants, namely the audience shot (ADS) and the musical instrument shot (MIS). The remaining 5 and 10 official concert videos, corresponding to 1,331 and 2,485 (high-resolution) music frames, are used for validation and testing.

4.1 Evaluation Protocols

In the experiments, we first report the performance of the default teacher network MF-RNNs (with film-language) [26] and its three assistant networks, which are trained with high, middle, and low temporal resolutions as hRNN, mRNN, and lRNN, respectively. Since our target is to obtain a high temporal resolution shot sequence, the translation results of mRNN and lRNN need up-sampling to match the number of frames needed by the high temporal resolution. To this end, we simply repeat the middle temporal resolution (or low temporal resolution) translations locally such that the final output has the same number of frames as the high temporal resolution. We then compare the performance of the MF-RNNs (with film-language) and a lightweight RNN trained by the conventional KD technique [8]. Finally, for the above competing methods, we show the performance of the proposed DIL mechanism, which includes assistant networks hRNN-DIL, mRNN-DIL, lRNN-DIL, teacher network MF-RNNs-DIL, and student network lightweight RNN-DIL. For the training of the above RNNs, we apply random initialization for the weights, a constant learning rate of 10^{-3} , the dropout to avoid over-fitting, and the RMSprop solver [7] for optimization. The meta-parameters of each method are set based on the validation error.

In the experiments, four metrics are used for performance evaluation, including the accuracy of shot category (ASC), the trend of shot change and the distance of shot type under the aligned category, which are named TSC-AC and DST-AC, respectively, and the duration of shot (DS) [24, 26]. Similar to the word accuracy widely used in speech recognition [29], the ASC is defined as

$$ASC = \frac{N - D - S - I}{N} \times 100\%, \quad (14)$$

where N is the number of categorized shots in the ground truth shot sequence. D, S , and I denote the deletion errors, substitution errors, and insertion errors of the shot category, respectively. According to observations, the visual storytelling of shots in the concert video relies mainly on shots defined by the language of film (see Table 1), supplemented by a small amount of the concert-specific shots (ADS and MIS). Therefore, in order to evaluate whether the concert-specific shots are correctly arranged in the language-of-film-based

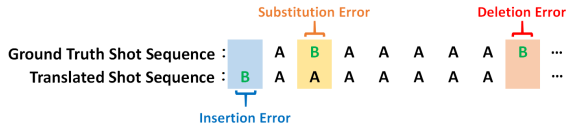


Figure 5: Illustration of the dynamic programming-derived shot category-aligned path for a ground truth shot sequence and a translated shot sequence.

shot sequence, for ASC measurement, the language-of-film-based shots and the concert-specific shots are first categorized into Category A and Category B, respectively. A dynamic programming algorithm [29] is employed to find the optimal alignment of shot categories between the ground truth shot sequence and the translated shot sequence, to obtain D , S , and I for ASC measurement, as shown in Figure 5. We further concentrate on the TSC-AC measurement in the aligned Category A and defined its metric as

$$TSC-AC = \frac{H}{M} \times 100\%, \quad (15)$$

where M is the total number of shot changes in the aligned Category A (*i.e.*, language-of-film-based shots), and H denotes the total number of correct shot change trends. According to the language of film, the types of shots are defined based on the viewing distance (from close view to far view, see Table 1). Therefore, two trends of shot changes are defined, including the increased viewing distance “+” and the decreased viewing distance “-”. For example, as shown in Figure 6, the shot type from MCU to LS is categorized into the trend of increased viewing distance “+”, while from XLS to MS is categorized into the trend of decreased viewing distance “-”. Thus, the aligned trends between the ground truth shot sequence and the translated shot sequence can be used to obtain H for evaluation of TSC-AC accuracy. To further validate the above-mentioned methods, the DST-AC metric is considered. We number the six types of shots (defined by language of film) as one to six according to the viewing distance. The numbered shots (see Figure 6) between aligned shot change trends are estimated one-by-one by setting DST-AC as the absolute error measurement. Finally, taking account of that the number of translated shots are different in each method, we report the unweighted average (better reflecting the imbalances among number of translated shots) time length of all eight shot types in the DS measurement. We report the average ASC, TSC-AC and DST-AC, and DS in the test set. In addition to the four evaluation metrics, we also show the number of parameters and model size for each competitive method as a cost comparison.

4.2 Results and Comparisons

Table 2 shows the performance of the ten methods for all the mentioned metrics. For comparison of the three temporal resolution RNNs, similar to the results in [26], the results first show that hRNN is superior to mRNN and IRNN in all the four metrics. According to observations, this could be due to that the use of an up-sampling technique in mRNN and IRNN could lose accuracy locally in translating the high temporal resolution. However, although hRNN outperformed mRNN and IRNN, the performance is still limited. One explanation for this may be that hRNN only considers adjacent memories, so it is difficult to handle the long-term dependencies over large and varied temporal intervals in the concert video. To

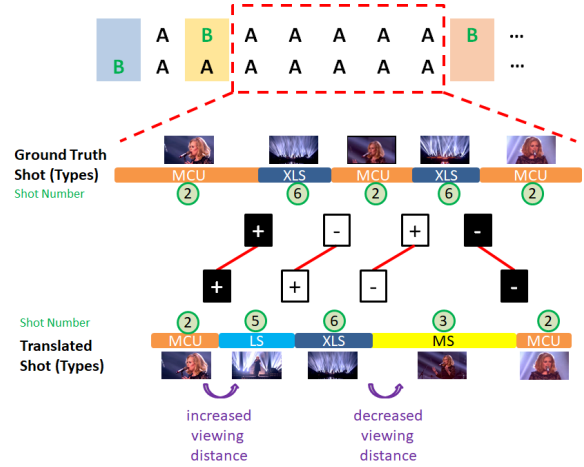


Figure 6: Illustration of the trends of shot changes under aligned Category A. Aligned black and white blocks indicate correct and incorrect trends for shot changes, respectively.

tackle the issue, MF-RNNs (with film-language) is proposed to introduce RNNs with various temporal resolutions to deal with such problem [26]. By appropriately integrating multi-resolution fused RNNs and film-language model, the results support that MF-RNNs (with film-language) can capture longer temporal dependencies and preserve long-term coherence among multiple shots (refer to ASC and TSC-AC metrics). That is, compared with hRNN, mRNN and IRNN, MF-RNNs (with film-language) can achieve more precise translation results. However, as mentioned earlier, for 1 second (high temporal resolution) music-to-shot translation, MF-RNNs (with film-language) requires reading music information for at least 4 seconds (low temporal resolution) and integrating multiple deep-net models. Such an ensemble framework makes MF-RNNs (with film-language) too heavy to deploy in practical applications. In addition, each temporal resolution RNN used in MF-RNNs (with film-language) is learned directly from ground truth (shot) label, which can only provide very limited information.

To make the network efficient and easy to deploy, KD [8] is applied to build a lightweight RNN (student network) that takes the music representation of high temporal resolution as input to mimic the output behavior of the teacher network MF-RNNs (with film-language). As can be seen from Table 2, the lightweight RNN-KD is comparable to MF-RNNs (with film-language), and sometimes even better (see ASC metric). A reasonable explanation is that the lightweight RNN-KD further considers the soft target distribution (distilled from teacher network) as supervision during the learning phase, which makes it easier to optimize and has better generalization capability for test data [3, 8, 12]. Moreover, a lightweight network structure with fewer parameters (see section 3.2 and Table 2) can also alleviate the phenomenon of overfitting. In addition, the results also show that the lightweight RNN-KD is superior to hRNN under the same network structure settings (see section 3.2 and Table 2). Such results confirm that by mimicking the output behavior (soft target distribution) of MF-RNNs (with film-language), lightweight RNN-KD can more effectively learn the long-term dependencies over large and varied temporal intervals and preserve the long-term coherence among multiple shots. However, despite

Table 2: Average ASC, TSC-AC and DST-AC as well as DS, number of parameters and model size.

Method \ Metric	hRNN [26] (Assistant Network)	mRNN [26] (Assistant Network)	IRNN [26] (Assistant Network)	MF-RNNs [26] (Teacher Network)	Lightweight RNN-KD [8] (Student Network)	hRNN-DIL (Assistant Network)	mRNN-DIL (Assistant Network)	IRNN-DIL (Assistant Network)	MF-RNNs-DIL (Teacher Network)	Lightweight RNN-DIL (Student Network)	Ground Truth
Avg. ASC (%)	64.21	62.15	51.58	67.61	70.11	65.44	64.31	55.87	72.63	71.63	-
Avg. TSC-AC (%)	65.93	58.00	53.37	70.90	69.64	73.39	64.16	58.87	74.75	74.88	-
Avg. DST-AC	2.01	2.09	2.17	1.50	1.72	1.90	1.90	2.07	1.49	1.69	-
DS (seconds)	3.12	4.40	5.97	3.32	3.22	3.15	4.19	5.57	3.02	3.49	3.50
#Parameters	1.75M	1.75M	1.75M	~5.25M	1.75M	1.75M	1.75M	1.75M	~5.25M	1.75M	-
Model Size (MB)	6.69	6.69	6.69	20.10	6.69	6.69	6.69	6.69	20.10	6.69	-

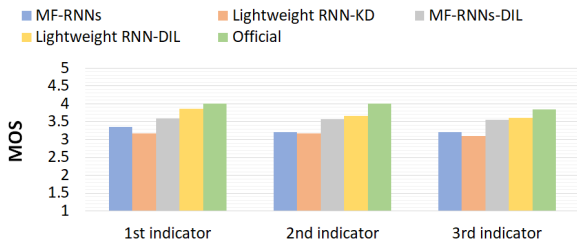


Figure 7: Results of the subjective MOS evaluation

the effectiveness of KD, such one-way knowledge transfer in KD limits the opportunity for the teacher network to grow with the student network, and thus indirectly harms the student network.

Based on the above analysis, we develop the DIL mechanism that enables teacher and student networks to learn and grow from each other. The results in Table 2 first show that under the DIL mechanism, the performance of the three assistant networks hRNN-DIL, mRNN-DIL, and IRNN-DIL is better than the original three, *i.e.*, hRMM, mRNN, and IRNN. Such results indicate that by using student feedback (soft target distribution) as supervision, the generalization capability of each assistant network can be improved. Once the assistant networks are upgraded, the performance of the teacher network (MF-RNNs-DIL) will be enhanced accordingly. Table 2 confirms the claim that MF-RNNs-DIL is superior to the original teacher network MF-RNNs (with film-language) in almost all metrics. Regarding the lightweight RNN comparison, it is evident from Table 2 that the proposed lightweight RNN-DIL outperforms lightweight RNN-KD in all metrics. It is even better than the original MF-RNNs (with film-language). The reason for this result is that under the DIL mechanism, the teacher network (MF-RNNs-DIL) can be upgraded based on the feedback from the student network (lightweight RNN-DIL), so it provides an opportunity for the student network to learn from a more powerful teacher network. Overall, the results indicate that under the DIL mechanism, the performance of both the lightweight RNN and the MF-RNNs (with film-language) can be further improved. Moreover, for music-to-shot translation, the lightweight-RNN-DIL only needs to read the music representation for one second (high temporal resolution), which is efficient.

Subjective evaluation in terms of 5-point mean opinion score (MOS) is conducted on three concert video sets¹. For a concert video set, each concert video is generated from multiple audience recordings that under the guidance of official shot type sequence and the shot type sequences translated from MF-RNNs (with film-language), lightweight RNN-KD, MF-RNNs-DIL, and lightweight RNN-DIL, respectively. For evaluation, we perform a blind test on each concert video set, which provides subjects with five concert videos in

¹The MOS results and three concert video sets are available at <https://sites.google.com/site/m2vsdil/mos>.

a random order. After viewing each concert video, the subject is asked to rate a MOS for three indicators: (1) Does the frequency of shot switching match the music? (2) Does the timing of cut point match the music? (3) Overall, does the visual storytelling of shots match the music? In total, each concert video is evaluated by 29 subjects (recruited from the authors’ laboratory and social media, aged between 20 and 66, with education from high school to PhD). The average MOS scores for the three indicators over all concert videos and subjects are shown in Figure 7. The results on the first and second indicators clearly show that the lightweight RNN-DIL can indeed generate more suitable shot sequence, which has the proper shot switching frequency and cut point timing for a music. Such results are reflected in objective evaluation (see Table 2), that is, the DS of the lightweight RNN-DIL is closest to the official (ground truth) concert video, so superior results can be obtained on both indicators. On the other hand, although MF-RNNs-DIL has better performance on ASC and DST-AC, due to the inaccuracy of DS, it limits the subjective experience in terms of these two indicators. However, the results show that the MF-RNNs-DIL is superior to original MF-RNNs (with film-language) and lightweight RNN-KD. Such results reveal that when evaluating the first and second indicators, the performance of ASC, TSC-AC, and DST-AC will also affect the subject’s perception. Finally, the average MOS score in the third indicator confirms that MF-RNNs-DIL and lightweight RNN-DIL perform well, and can generate appealing music-compliant professional-like concert videos with better viewing and listening experiences. We noticed that the average MOS scores of lightweight RNN-DIL and MF-RNNs-DIL are also quite close to the official concert video, which is really encouraging. The video demo can be found at <https://sites.google.com/site/m2vsdil/demo>.

5 CONCLUSIONS

This study introduces a novel deep interactive learning (DIL) mechanism that enables the interactive transfer of knowledge between an ensemble teacher network and a student network to build a compact yet accurate sequence-to-sequence model for music to visual storytelling of shots translation. Experiments on both objective and subjective evaluations demonstrate that the DIL mechanism outperforms the competing methods, and achieves the state-of-the-art results by using a much smaller network. Leveraging with these promising outcomes, our future work along this line would focus on addressing the challenging issues of learning from music to generate visual effects, which is essential to increase the quality of audiovisual experience.

ACKNOWLEDGMENTS

This work was supported in part by MOST grants 109-2634-F-001-011, 109-2634-F-001-012, 107-2218-E-001-010-MY2, 107-2628-E-009-007-MY3, 109-2634-F-007-013, and 109-2221-E-009-113-MY3.

REFERENCES

- [1] D. Andrews. 2011. *Digital overdrive: Communications & multimedia technology*. Digital Overdrive.
- [2] J. Ba and R. Caruana. 2014. Do deep nets really need to be deep?. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2654–2662.
- [3] C. Bucilă, R. Caruana, and A. Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 535–541.
- [4] P. Chaudhar, A. Choromansk, S. Soatt, Y. LeCun, C. Baldass, C. Borg, J. Chays, L. Sagun, and R. Zecchina. 2017. Entropy-SGD: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations (ICLR)*.
- [5] S. Han, H. Mao, and W. J. Dally. 2016. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *International Conference on Learning Representations (ICLR)*.
- [6] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems (NeurIPS)*. 820–828.
- [7] G. E. Hinton, N. Srivastava, and K. Swersky. 2012. Lecture 6a overview of mini-batch gradient descent. *Coursera Lecture Slides* (2012). [online] Available: <https://NeurIPSclass.coursera.org/neuralnets-2012-001/lecture>.
- [8] G. E. Hinton, O. Vinyals, and J. Dean. 2015. Distilling the knowledge in a neural network. In *NeurIPS Workshop*.
- [9] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf. 2017. Pruning filters for efficient ConvNets. In *International Conference on Learning Representations (ICLR)*.
- [10] J.-C. Lin, W.-L. Wei, T.-L. Liu, Y.-H. Yang, H.-M. Wang, H.-R. Tyan, and H.-Y. Mark Liao. 2018. Coherent deep-net fusion to classify shots in concert videos. *IEEE Transactions on Multimedia* 20, 11 (2018), 3123–3136.
- [11] J.-Y. Liu and Y.-H. Yang. 2016. Event localization in music auto-tagging. In *Proc. ACM Multimedia*. 1048–1057.
- [12] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang. 2019. Structured knowledge distillation for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2604–2613.
- [13] Y. Liu, J. Fu, T. Mei, and C. W. Chen. 2017. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *AAAI Conference on Artificial Intelligence*.
- [14] Y. Liu, L. Sheng, J. Shao, J. Yan, S. Xiang, and C. Pan. 2018. Multi-label image classification via knowledge distillation from weakly-supervised detection. In *Proc. ACM Multimedia*. 700–708.
- [15] G. Mercado. 2010. *The filmmaker's eye: Learning (and breaking) the rules of cinematic composition*. Taylor & Francis.
- [16] H. Pan, Liang Z.-P., and T. S. Huang. 2001. Estimation of the joint probability of multisensory signals. *Pattern Recognition Letters* 22, 13 (2001), 1431–1437.
- [17] C. C. Park and G. Kim. 2015. Expressing an image stream with a sequence of natural sentences. In *Advances in Neural Information Processing Systems (NeurIPS)*. 73–81.
- [18] G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, and G. E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *ICLR Workshop*.
- [19] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. 2016. XNOR-net: ImageNet classification using binary convolutional neural networks. In *European Conference on Computer Vision (ECCV)*.
- [20] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. 2014. FitNets: Hints for thin deep nets. *International Conference on Learning Representations (ICLR)*.
- [21] M. K. Saini, R. Gadde, S. Yan, and W.T. Ooi. 2012. MoViMash: online mobile video mashup. In *Proc. ACM Multimedia*. 139–148.
- [22] B. B. Sau and V. N. Balasubramanian. 2016. Deep model compression: Distilling knowledge from noisy teachers. *arXiv:1610.09650* (2016).
- [23] Y. Shi, M.-Y. Hwang, X. Lei, and H. Sheng. 2019. Knowledge distillation for recurrent neural network language modeling with trust regularization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7230–7234.
- [24] W.-L. Wei, J.-C. Lin, T.-L. Liu, H.-R. Tyan, H.-M. Wang, and H.-Y. Mark Liao. 2020. Learning to visualize music through shot sequence for automatic concert video mashup. *to appear in IEEE Transactions on Multimedia* (2020).
- [25] W.-L. Wei, J.-C. Lin, T.-L. Liu, Y.-H. Yang, H.-M. Wang, H.-R. Tyan, and H.-Y. Mark Liao. 2017. Deep-net fusion to classify shots in concert videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1383–1387.
- [26] W.-L. Wei, J.-C. Lin, T.-L. Liu, Y.-H. Yang, H.-M. Wang, H.-R. Tyan, and H.-Y. Mark Liao. 2018. Seethevoice: Learning from music to visual storytelling of shots. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*. 1–6.
- [27] H.-Y. Wu, F. Palù, R. Ranon, and M. Christie. 2018. Thinking like a director: film editing patterns for virtual cinematographic storytelling. *ACM Transactions on Multimedia Computing, Communications and Applications* 14, 4 (2018), 1–23.
- [28] J. Yim, D. Joo, J. Bae, and J. Kim. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4133–4141.
- [29] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. 2006. *The HTK Book Version 3.4*. Cambridge Univ. Press.
- [30] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng. 2019. Revisit knowledge distillation: a teacher-free framework. *ArXiv abs/1909.11723* (2019).
- [31] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. 2018. Deep mutual learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4320–4328.