# DIFFICULTY-AWARE NEURAL BAND-TO-PIANO SCORE ARRANGEMENT BASED ON NOTE- AND STATISTIC-LEVEL CRITERIA

*Moyu Terao*[1] *Yuki Hiramatsu*[1] *Ryoto Ishizuka*[1] *Yiming Wu*[1] *Kazuyoshi Yoshii*[1,2]

[1]Graduate School of Informatics, Kyoto University, Japan
[2]PRESTO, Japan Science and Technology Agency (JST), Japan

## ABSTRACT

This paper describes a neural music arrangement method that converts a given band score into a piano score with an elementary or advanced level. The major challenge of this task lies in its ill-posed nature, *i.e.*, various piano arrangements are plausible for a band score. In this paper, we take a score reduction approach based on supervised training of a mask estimation network (U-Net) with note- and statistic-level criteria. Based on statistical analysis of existing piano arrangements, a reasonable piano score is assumed to be obtained by reducing an *augmented* band score obtained by up- and down-shifting an original band score by one octave. This effectively narrows down a solution space. At the heart of our approach is to train a U-Net conditioned by a given difficulty level such that a piano score obtained by masking an augmented band score is close to the ground-truth piano score not only at a *note* level but also at a *statistic* level. We focus on three kinds of note statistics, *i.e.*, a distribution of the numbers of concurrent notes, that of the intervals between the highest and lowest pitches, and that of the per-measure numbers of notes. The experimental results show the importance of both the instance- and meta-level criteria for supervised training.

***Index Terms—*** Automatic piano arrangement, score reduction, symbolic music processing, deep learning

## 1. INTRODUCTION

Music arrangement refers to changing the instrumentation of a musical piece while preserving the content. Much effort has been devoted to automatic piano arrangement [1–5], guitar arrangement [6–8], and orchestration [9, 10]. In recent years, deep neural networks (DNNs) have intensively been used for automatic piano arrangement because of their rich expression capability required for music [11–13].

The fundamental problem of automatic piano arrangement lies in its ill-posed nature, *i.e.*, the "ground-truth" arrangement cannot be uniquely determined for a given musical piece. Nonetheless, we need to find the best arrangement that preserves the original score characteristics. One thus might train a score-to-score (*e.g.*, band-to-piano) conversion network in a supervised manner, where only one of the infinitely many plausible piano scores is given as the ground truth with a particular difficulty level. Such a network is typically optimized such that the note-level matching between the estimated and ground-truth scores is maximized. The possibility that the estimated piano score is an alternative ground-truth score, however, is not considered. A score-to-score mapping is thus hard to learn stably in the standard one-to-one supervised training.

To tackle the ill-posed problem, we propose a score reduction approach to difficulty-aware band-to-piano score arrangement that aims to yield a *locally* and *globally* coherent piano score for a given band score (Fig. 1). Our approach effectively narrows down a solution space to preserve the characteristics of a band score. Prior to the method design, we investigated the relationships between the left- and right-hand parts of piano scores and the melody and accompaniment parts of the corresponding band scores. Based on this statistical analysis, we assume that a reasonable piano score can be obtained by reducing an *augmented band score* obtained by up- and down-shifting an original band score by one octave. We also assume that a piano score is an instance drawn from a certain probabilistic arrangement model and the statistics of piano notes are thus less sensitive to the preferences of arrangers than the concrete note arrangements. As note statistics that strongly reflect the performance difficulty, we focus on a distribution of polyphony levels (numbers of concurrent notes), that of polyphony widths (intervals between the highest and lowest pitches), and that of note densities (per-measure numbers of notes). A mask estimation network is trained in a regularized manner, conditioned by a given difficulty level, such that a piano score obtained by masking an augmented band score is close to the ground-truth piano score not only at a *note* level (instance level) but also at a *statistic* level (meta level).

The main contribution of this study is to propose a new strategy of supervised piano arrangement based on the local and global evaluation of piano scores to deal with the non-uniqueness (ill-posedness) of piano arrangement. The proposed regularization of the statistic-level note statistics plays an essential role, especially when only a limited amount of paired data with a limited variety are available for supervised training. We experimentally show the performance improvement of piano arrangement.

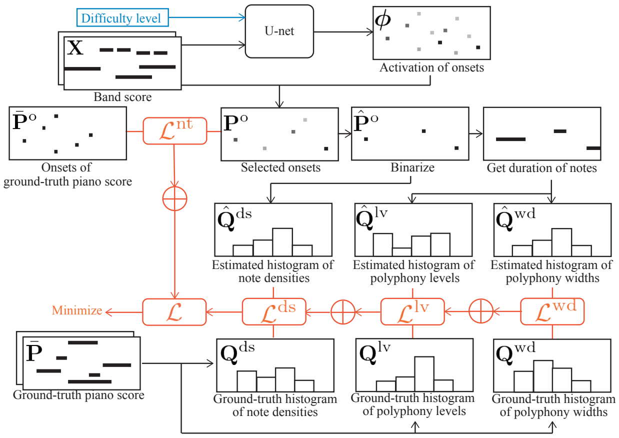
Difficulty level
U-net
$\phi$
Activation of onsets
X
Band score
$\bar{\mathbf{P}}^{\mathrm{o}}$
Onsets of ground-truth piano score
$\mathcal{L}^{\mathrm{nt}}$
$\mathbf{P}^{\mathrm{o}}$
Selected onsets
$\hat{\mathbf{P}}^{\mathrm{o}}$
Binarize
Get duration of notes
$\hat{\mathbf{Q}}^{\mathrm{ds}}$
Estimated histogram of note densities
$\hat{\mathbf{Q}}^{\mathrm{lv}}$
Estimated histogram of polyphony levels
$\hat{\mathbf{Q}}^{\mathrm{wd}}$
Estimated histogram of polyphony widths
Minimize
$\mathcal{L}$
$\mathcal{L}^{\mathrm{ds}}$
$\mathcal{L}^{\mathrm{lv}}$
$\mathcal{L}^{\mathrm{wd}}$
$\bar{\mathbf{P}}$
Ground-truth piano score
$\mathbf{Q}^{\mathrm{ds}}$
Ground-truth histogram of note densities
$\mathbf{Q}^{\mathrm{lv}}$
Ground-truth histogram of polyphony levels
$\mathbf{Q}^{\mathrm{wd}}$
Ground-truth histogram of polyphony widths

**Fig. 1**. Our score reduction approach to band-to-piano score arrangement based on supervised training of a mask estimation network with a statistic-level regularization mechanism.

This work is supported in part by JST PRESTO No. JPMJPR20CB and JSPS KAKENHI Nos. 19H04137 and 20K21813.

## 2. RELATED WORK

The playable measure has been proposed for judging whether a piano score can actually be performed by humans and is the basis of defining the performance difficulty. Chiu *et al.* [1], for example, proposed a score reduction method that selects important musical phrases in terms of utility and playability such that 1) at most five phrases are played simultaneously and that 2) the interval between the highest and lowest pitches of concurrent notes are within the size of each hand. Onuma *et al.* [2] found that such playability conditions are often violated through investigation of problem-solving operations in music arrangements made by human composers. Nakamura *et al.* [5] extended a score reduction method based on a probabilistic model of a piano score [14] to incorporate fingering estimated by an HMM [15, 16]. Considering that the playability depends on the player's skill, they attempted to quantify the performance difficulty. It has been shown that the polyphony levels and widths of a piano score are closely related to the validity and difficulty of the score. Wang *et al.* [11] proposed a DNN-based piano arrangement method using a popular music dataset called Pop909 including multiple piano arrangements with beat, chords, and key annotations. They investigated polyphonic music generation based on the transformer [17] and accompaniment generation from melodies.

## 3. PROPOSED METHOD

This section describes the proposed method that arranges a band score into a piano score with a specified difficulty level using the note- and statistic-level criteria.

### 3.1. Problem Specification

Our goal is to convert a band score $\mathbf{B} \triangleq \{\mathbf{B}_{\mathrm{A}}, \mathbf{B}_{\mathrm{M}}\}$ into a piano score $\hat{\mathbf{P}} \triangleq \{\hat{\mathbf{P}}_{\mathrm{L}}, \hat{\mathbf{P}}_{\mathrm{R}}\}$ with a specified difficulty level $L \in \{0, 1\}$, where $\mathbf{B}_{\mathrm{A}} \triangleq \{\mathbf{B}^{\mathrm{o}}_{\mathrm{A}}, \mathbf{B}^{\mathrm{p}}_{\mathrm{A}}\}$ and $\mathbf{B}_{\mathrm{M}} \triangleq \{\mathbf{B}^{\mathrm{o}}_{\mathrm{M}}, \mathbf{B}^{\mathrm{p}}_{\mathrm{M}}\}$ denote the accompaniment and melody parts of the band score, respectively, $\hat{\mathbf{P}}_{\mathrm{L}} \triangleq \{\hat{\mathbf{P}}^{\mathrm{o}}_{\mathrm{L}}, \hat{\mathbf{P}}^{\mathrm{p}}_{\mathrm{L}}\}$ and $\hat{\mathbf{P}}_{\mathrm{R}} \triangleq \{\hat{\mathbf{P}}^{\mathrm{o}}_{\mathrm{R}}, \hat{\mathbf{P}}^{\mathrm{p}}_{\mathrm{R}}\}$ are the left- and right-hand parts of the piano score, respectively, and $L = 0$ and $L = 1$ denote the elementary and advanced levels, respectively. Each of these four parts is represented by a pair of an onset matrix and a pitch matrix of size $P \times N$ (denoted by $*^{\mathrm{o}}$ and $*^{\mathrm{p}}$, respectively), where $P$ denotes the number of pitches (MIDI note numbers) and $N$ denotes the number of tatums in 16th-note units ($P = 128$ and $N = 16 \times 12 = 192$ in this paper). For example, $\mathbf{B}^{\mathrm{o}}_{\mathrm{M}}(p, n) = 1$ denotes the presence of an onset at pitch $p$ and tatum $n$ and $\mathbf{B}^{\mathrm{p}}_{\mathrm{M}}(p, n) = 1$ denotes the presence of pitch at pitch $p$ at tatum $n$. Let $h \in \{\mathrm{L}, \mathrm{R}\}$ denote a left- or right-hand part.

### 3.2. Score Reduction Approach

Based on the statistical analysis of existing piano arrangements (see details in Section 4.2), we assume that a reasonable piano score can be obtained as a subset of an *augmented* band score. Let $\mathbf{Z}^{\mathrm{o}} \in \{0, 1\}^{P \times N}$ be an *augmented* onset matrix given by

$$\mathbf{Z}^{\mathrm{o}}(p, n) = \max_{j \in \{-12, 0, 12\}} \left(\mathbf{B}^{\mathrm{o}}_{\mathrm{A}}(p + j, n), \mathbf{B}^{\mathrm{o}}_{\mathrm{M}}(p + j, n)\right). \qquad (1)$$

Let $\boldsymbol{\phi} \triangleq \{\boldsymbol{\phi}_{\mathrm{L}}, \boldsymbol{\phi}_{\mathrm{R}}\}$ be a pair of soft mask matrices $\boldsymbol{\phi}_{\mathrm{L}} \in [0, 1]^{P \times N}$ and $\boldsymbol{\phi}_{\mathrm{R}} \in [0, 1]^{P \times N}$ for the left- and right-hand parts, respectively, which are estimated by a DNN (see details in Section 3.3). Let $\mathbf{P}^{\mathrm{o}} \triangleq \{\mathbf{P}^{\mathrm{o}}_{\mathrm{L}}, \mathbf{P}^{\mathrm{o}}_{\mathrm{R}}\}$ be a pair of onset probability matrices $\mathbf{P}^{\mathrm{o}}_{\mathrm{L}}, \mathbf{P}^{\mathrm{o}}_{\mathrm{R}} \in [0, 1]^{P \times N}$ for the left- and right-hand parts, respectively, which are given by a soft masking process as follows:

$$\mathbf{P}^{\mathrm{o}}_{\mathrm{L}} = \boldsymbol{\phi}_{\mathrm{L}} \odot \mathbf{Z}^{\mathrm{o}}, \quad \mathbf{P}^{\mathrm{o}}_{\mathrm{R}} = \boldsymbol{\phi}_{\mathrm{R}} \odot \mathbf{Z}^{\mathrm{o}}, \qquad (2)$$

where $\odot$ denotes the element-wise product. Finally, the pair of the onset matrices $\hat{\mathbf{P}}^{\mathrm{o}} \triangleq \{\hat{\mathbf{P}}^{\mathrm{o}}_{\mathrm{L}}, \hat{\mathbf{P}}^{\mathrm{o}}_{\mathrm{R}}\}$ are obtained by binarizing $\mathbf{P}^{\mathrm{o}} \triangleq \{\mathbf{P}^{\mathrm{o}}_{\mathrm{L}}, \mathbf{P}^{\mathrm{o}}_{\mathrm{R}}\}$ with a threshold (0.5 in this paper). The pair of the pitch matrices $\hat{\mathbf{P}}^{\mathrm{p}} \triangleq \{\hat{\mathbf{P}}^{\mathrm{p}}_{\mathrm{L}}, \hat{\mathbf{P}}^{\mathrm{p}}_{\mathrm{R}}\}$ are derived from the pair of the pitch matrices $\mathbf{B}^{\mathrm{p}} \triangleq \{\mathbf{B}^{\mathrm{p}}_{\mathrm{L}}, \mathbf{B}^{\mathrm{p}}_{\mathrm{R}}\}$ of the band score according to the estimated $\hat{\mathbf{P}}^{\mathrm{o}}$ (see details in Section 3.3.2).

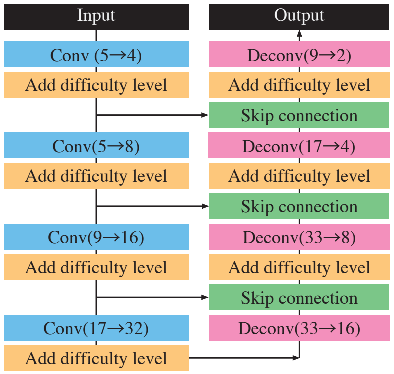

**Fig. 2**. The architecture of U-Net. The number of channels is gradually changed as shown in parentheses.

### 3.3. Onset Mask Estimation

The soft mask $\boldsymbol{\phi}$ is estimated with a U-Net [18] that takes as input a band score $\mathbf{B}$ and a difficulty level $L$ (Fig. 2). Specifically, we make a tensor $\mathbf{X} \triangleq \{\mathbf{B}^{\mathrm{o}}_{\mathrm{A}}, \mathbf{B}^{\mathrm{p}}_{\mathrm{A}}, \mathbf{B}^{\mathrm{o}}_{\mathrm{M}}, \mathbf{B}^{\mathrm{p}}_{\mathrm{M}}, \mathbf{L}\} \in \{0, 1\}^{5 \times P \times N}$, where $\mathbf{L} \in \{0, 1\}^{P \times N}$ is a matrix whose elements are all equal to the difficulty level $L$. The output is given by a sigmoid function to limit the values of $\boldsymbol{\phi}$ in $[0, 1]$. Given a number of pairs of band scores and the corresponding piano scores, the U-Net is trained in a supervised manner such that a weighted sum of note- and statistic-level losses that evaluate the output $\boldsymbol{\phi}$ is minimized. First, the estimated piano score should be close to the ground-truth piano score. Second, the note statistics of the estimated piano score should be close to those of existing piano scores with a specified difficulty level. Thus, the loss function consists of *note-level* and *statistic-level* losses.

#### 3.3.1. Note-Level Loss

The note-level loss is defined as a modified version of the cross entropy loss as follows:

$$\mathcal{L}^{\mathrm{nt}} = -\sum_{h \in \{\mathrm{L}, \mathrm{R}\}} \sum_{p=1}^{P} \sum_{n=1}^{N} \Big(\alpha \cdot \bar{\mathbf{P}}^{\mathrm{o}}_{h}(p, n) \log \mathbf{P}^{\mathrm{o}}_{h}(p, n) + \big(1 - \bar{\mathbf{P}}^{\mathrm{o}}_{h}(p, n)\big) \log\big(1 - \mathbf{P}^{\mathrm{o}}_{h}(p, n)\big)\Big), \qquad (3)$$

where $\bar{\mathbf{P}}^{\mathrm{o}} \triangleq \{\bar{\mathbf{P}}^{\mathrm{o}}_{\mathrm{L}}, \bar{\mathbf{P}}^{\mathrm{o}}_{\mathrm{R}}\}$ represents a pair of the ground-truth onset matrices $\bar{\mathbf{P}}^{\mathrm{o}}_{\mathrm{L}}, \bar{\mathbf{P}}^{\mathrm{o}}_{\mathrm{R}} \in \{0, 1\}^{P \times N}$ for the left- and right-hand parts and $\alpha \geq 0$ is a weighting factor used for compensating for the imbalance of the numbers of onset and non-onset frames ($\alpha = 4$ in this paper).

#### 3.3.2. Statistic-Level Losses

Using the Gumbel-sigmoid trick [19], the pair of the onset matrices $\hat{\mathbf{P}}^{\mathrm{o}} \triangleq \{\hat{\mathbf{P}}^{\mathrm{o}}_{\mathrm{L}}, \hat{\mathbf{P}}^{\mathrm{o}}_{\mathrm{R}}\}$ is stochastically drawn from the pair of the onset probability matrices $\mathbf{P}^{\mathrm{o}} \triangleq \{\mathbf{P}^{\mathrm{o}}_{\mathrm{L}}, \mathbf{P}^{\mathrm{o}}_{\mathrm{R}}\}$ given by Eq. (2) in a differentiable manner. We here aim to compute the pair of the pitch matrices $\hat{\mathbf{P}}^{\mathrm{p}} = \{\hat{\mathbf{P}}^{\mathrm{p}}_{\mathrm{L}}, \hat{\mathbf{P}}^{\mathrm{p}}_{\mathrm{R}}\}$ from $\hat{\mathbf{P}}^{\mathrm{o}}$ and the augmented band score (Fig. 3). The augmented onset matrix $\mathbf{Z}^{\mathrm{o}}$ is first converted into a sequence of pitch-onset pairs $\{(p_a, n_a)\}_{a=1}^{A}$, where $A$ is the number of onsets in the augmented band score, and $p_a$ and $n_a$ represent the pitch and onset tatum of note $a$, respectively, *i.e.*, the number of $(p, n)$'s such that $\mathbf{Z}^{\mathrm{o}}(p, n) = 1$ is equal to $A$. Let $\mathbf{A} \triangleq \{p_a\}_{a=1}^{A}$ be a pitch sequence of the augmented band score. Let $\tilde{\mathbf{Z}} \triangleq \{\tilde{\mathbf{Z}}^{\mathrm{o}}, \tilde{\mathbf{Z}}^{\mathrm{p}}\}$ be a pair of onset and pitch matrices $\tilde{\mathbf{Z}}^{\mathrm{o}}, \tilde{\mathbf{Z}}^{\mathrm{p}} \in \{0,1\}^{A\times N}$ determined by $\mathbf{Z}^{\mathrm{o}}$, where $\tilde{\mathbf{Z}}^{\mathrm{o}}(a,n) = 1$ when $n_a = n$, and $\tilde{\mathbf{Z}}^{\mathrm{p}}(a,n) = 1$ when note $a$ is activated at tatum $n$. Let $\mathbf{E} \triangleq \{\mathbf{E}_{\mathrm{L}}, \mathbf{E}_{\mathrm{R}}\}$ be the pair of onset matrices $\mathbf{E}_{\mathrm{L}}, \mathbf{E}_{\mathrm{R}} \in \{0,1\}^{A\times N}$ for the left- and right-hand parts determined by $\hat{\mathbf{P}}^{\mathrm{o}}$, where the $a$-th row of $\mathbf{E}_h$ is given by the $p_a$-th row of $\hat{\mathbf{P}}^{\mathrm{o}}_h$, *e.g.*, $\mathbf{E}_h(a) = \hat{\mathbf{P}}^{\mathrm{o}}_h(p_a)$. The element-wise product of $\mathbf{E}_h$ and $\tilde{\mathbf{Z}}^{\mathrm{o}}$ is then accumulated along the row direction, resulting in $\mathbf{S} \triangleq \{\mathbf{S}_{\mathrm{L}}, \mathbf{S}_{\mathrm{R}}\} \in \{0,1\}^{2\times A}$, where $\mathbf{S}_{\mathrm{L}}(a) = 1$ or $\mathbf{S}_{\mathrm{R}}(a) = 1$ represents the presence of note $a$ in the left- or right-hand part of the piano score. Note duration matrix $\mathbf{D} \triangleq \{\mathbf{D}_{\mathrm{L}}, \mathbf{D}_{\mathrm{R}}\} \in \{0,1\}^{2\times A\times N}$ is also obtained by multiplying $\mathbf{S}$ and $\tilde{\mathbf{Z}}^{\mathrm{p}}$. Finally, the pair of the pitch matrices $\hat{\mathbf{P}}^{\mathrm{p}} = \{\hat{\mathbf{P}}^{\mathrm{p}}_{\mathrm{L}}, \hat{\mathbf{P}}^{\mathrm{p}}_{\mathrm{R}}\} \in \{0,1\}^{2\times P\times N}$ is obtained by sorting the notes of $\mathbf{D}$ in the pitch-ascending order.

At the heart of our method is to compute the distributions (histograms) of the three kinds of note statistics in each mini-batch in a *differentiable* manner for regularizing the U-Net. Let $\hat{\mathbf{C}}^{\mathrm{lv}}_h(n)$, $\hat{\mathbf{C}}^{\mathrm{wd}}_h(n)$, and $\hat{\mathbf{C}}^{\mathrm{ds}}_h(m)$ denote the polyphony level (the number of concurrent pitches) at tatum $n$, the polyphony width (the interval between the highest and lowest pitches) at tatum $n$, and the note density (the number of notes) at measure $m$, respectively, in the estimated piano score, which are given by

$$\hat{\mathbf{C}}^{\mathrm{lv}}_h(n) = \sum_{p=1}^{P} \hat{\mathbf{P}}^{\mathrm{p}}_h(p,n), \tag{4}$$

$$\hat{\mathbf{C}}^{\mathrm{wd}}_h(n) = \max_{a\in J}(\mathbf{D}_h(n) \odot \mathbf{A}^{+})_a - \min_{a\in J}(\mathbf{D}_h(n) \odot \mathbf{A}^{+})_a, \tag{5}$$

$$\hat{\mathbf{C}}^{\mathrm{ds}}_h(m) = \sum_{p=1}^{P}\sum_{n=1}^{16} \hat{\mathbf{P}}^{\mathrm{o}}_h(p, 16m+n), \tag{6}$$

where $\mathbf{A}^{+} \triangleq \{p^{+}_a\}_{a=1}^{A}$ represents a pitch sequence such that $p^{+}_a = p_a + 1$, $\mathbf{D}_h(n)$ represents the $n$-th column of $\mathbf{D}_h$, and $J \triangleq \{a \mid (\mathbf{D}_h(n) \odot \mathbf{A}^{+})_a > 0\}$.

We explain how to calculate the histogram and statistic-level loss for the polyphony level because the others are calculated in the same way. For each polyphony level $i$ $(0 \le i \le I)$, we first compute an auxiliary value $\mathbf{G}^{\mathrm{lv}}_h(i)$ accumulated over all tatums as follows:

$$\hat{\mathbf{G}}^{\mathrm{lv}}_h(i) = \sum_{n=1}^{N} \mathrm{ReLU}\left(-\hat{\mathbf{C}}^{\mathrm{lv}}_h(n) + i\right), \tag{7}$$

where $\mathrm{ReLU}(-\hat{\mathbf{C}}^{\mathrm{lv}}_h(n)+i)$ takes a positive number when $\hat{\mathbf{C}}^{\mathrm{lv}}_h(n) < i$ and takes zero otherwise, and $I$ is the maximum polyphony level. Note that $\hat{\mathbf{G}}^{\mathrm{lv}}_h(i)$ can be represented in a different way as follows:

$$\hat{\mathbf{G}}^{\mathrm{lv}}_h(i) = \sum_{j=0}^{i-1} \hat{\mathbf{F}}^{\mathrm{lv}}_h(j) \times (i-j), \tag{8}$$

where $\mathbf{F}^{\mathrm{lv}}_h(j)$ represents the frequency of polyphony level $j$ over all tatums. Thus, $\mathbf{F}^{\mathrm{lv}}_h \in [0,1]^{I+1}$ can be computed recursively as follows:

$$\hat{\mathbf{F}}^{\mathrm{lv}}_h(0) = \hat{\mathbf{G}}^{\mathrm{lv}}_h(0), \tag{9}$$

$$\hat{\mathbf{F}}^{\mathrm{lv}}_h(i) = \hat{\mathbf{G}}^{\mathrm{lv}}_h(i+1) - 2\hat{\mathbf{G}}^{\mathrm{lv}}_h(i) + \hat{\mathbf{G}}^{\mathrm{lv}}_h(i-1) \quad (i \ge 1). \tag{10}$$

The distribution of polyphony levels, denoted by $\hat{\mathbf{Q}}^{\mathrm{lv}}_h \in [0,1]^{I+1}$, is obtained by normalizing the frequencies $\{\hat{\mathbf{F}}^{\mathrm{lv}}_h(i)\}_{i=0}^{I}$ over all possible polyphony levels. The distributions of note widths and densities, denoted by $\hat{\mathbf{Q}}^{\mathrm{wd}}_h \in [0,1]^{J+1}$ and $\hat{\mathbf{Q}}^{\mathrm{ds}}_h \in [0,1]^{K+1}$, respectively, are computed in the same way, where $J$ and $K$ represent the maximum note width and the maximum number of notes, respectively, except that the accumulation operation is performed over measures instead of tatums in computing $\hat{\mathbf{Q}}^{\mathrm{ds}}_h$.

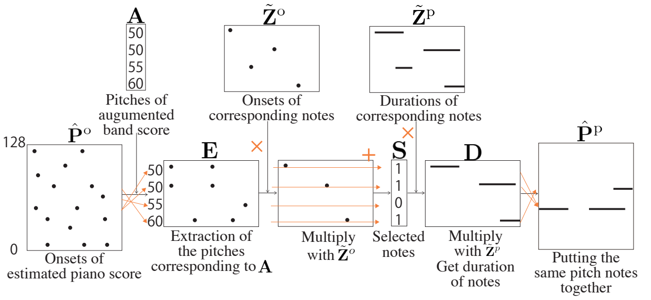
A
50
50
55
60
Pitches of augumented band score
$\tilde{\mathbf{Z}}^{\mathrm{o}}$
Onsets of corresponding notes
$\tilde{\mathbf{Z}}^{\mathrm{p}}$
Durations of corresponding notes
$\hat{\mathbf{P}}^{\mathrm{o}}$
128
0
Onsets of estimated piano score
E
×
50
50
55
60
Extraction of the pitches corresponding to A
+
Multiply with $\mathbf{Z}^o$
×
S
1
1
0
1
Selected notes
D
Multiply with $\mathbf{Z}^p$
Get duration of notes
$\hat{\mathbf{P}}^{\mathrm{p}}$
Putting the same pitch notes together

**Fig. 3**. Computation of the pitch matrices $\hat{\mathbf{P}}^{\mathrm{p}} = \{\hat{\mathbf{P}}^{\mathrm{p}}_{\mathrm{L}}, \hat{\mathbf{P}}^{\mathrm{p}}_{\mathrm{R}}\}$ for the left- and right-hand parts.

The statistic-level loss is measured based on the Jensen-Shannon (JS) divergence $\mathcal{D}_{\mathrm{JS}}$ between the ground-truth distribution $\mathbf{Q}^{*}_h$ and the estimated distribution $\hat{\mathbf{Q}}^{*}_h$ as follows:

$$\mathcal{L}^{*} = \sum_{h\in\{\mathrm{L},\mathrm{R}\}} \mathcal{D}_{\mathrm{JS}}\left(\mathbf{Q}^{*}_h \parallel \hat{\mathbf{Q}}^{*}_h\right) \quad (* \in \{\mathrm{lv}, \mathrm{wd}, \mathrm{ds}\}). \tag{11}$$

The total loss function $\mathcal{L}$ used for training the U-Net is given by

$$\mathcal{L} = \beta^{\mathrm{nt}}\mathcal{L}^{\mathrm{nt}} + \beta^{\mathrm{lv}}\mathcal{L}^{\mathrm{lv}} + \beta^{\mathrm{wd}}\mathcal{L}^{\mathrm{wd}} + \beta^{\mathrm{ds}}\mathcal{L}^{\mathrm{ds}}, \tag{12}$$

where $\beta^{\mathrm{nt}}$, $\beta^{\mathrm{lv}}$, $\beta^{\mathrm{wd}}$, and $\beta^{\mathrm{ds}}$ are weighting factors (hyperparameters) corresponding to the note-level loss and the thee statistic-level losses, respectively.

## 4. EVALUATION

This section reports two experiments We first evaluated the validity of the score reduction approach. We then examined the performance of the proposed method in terms of note- and statistic-level agreements between estimated and ground-truth piano scores.

### 4.1. Experimental Conditions

We collected 184 pairs of band and piano scores consisting of 85 pairs with the elementary level and 99 pairs with the advanced level. We conducted four-fold cross-validation. The number of measures was fixed to 12 and the time signature was assumed to be 4/4, *i.e.*, $N = 16\times 12 = 192$. For songs with other time signatures, in a measure shorter than sixteen tatums, the rest of the measure was silent. Otherwise, only the first sixteen tatums were used. We transposed the training data by one through eleven semitones for data augmentation. The U-Net is consisted of four convolutional layers and four deconvolutional layers. In each layer, $\mathbf{L}$ was stacked and batch regularization was performed. The kernel size was set to 4, the stride was set to 2, and the padding was set to 1. Dropout with a probability of 0.5 was applied to all the deconvolutional layers to prevent overfitting. Adam [20] with a learning rate of $10^{-4}$ was used to optimize the U-Net. The weighting factor was set to $\alpha = 4$. We adopted a threshold that maximizes the $\mathcal{F}$ for the right and left hands, for the validation data. Unless otherwise noted, the weighting factors were set to $\beta^{\mathrm{nt}} = 1$, $\beta^{\mathrm{lv}} = 0.4$, $\beta^{\mathrm{wd}}$, and $\beta^{\mathrm{ds}} = 0.1$ based on grid search on validation data. When optimising with $\mathcal{L}$, the weighting factor of the note-level loss was set to $\beta^{\mathrm{nt}} = 10$. We compared the proposed method minimizing the total loss $\mathcal{L}$ with a baseline method minimizing the note-level loss $\mathcal{L}^{\mathrm{nt}}$ only and an ablated method minimizing a weighted sum of $\mathcal{L}^{\mathrm{nt}}$ and one of the statistic-level losses $\mathcal{L}^{\mathrm{lv}}$, $\mathcal{L}^{\mathrm{wd}}$, and $\mathcal{L}^{\mathrm{ds}}$.

### 4.2. Experimental Results

The origins of the left- and right-hand notes of existing piano arrangements are shown in Fig. 4. We found that 76% of the right-hand notes and 72% of the left-hand notes in piano scores were derived from original band scores. This indicates that it is hard to make an

**Table 1**. Experimental results.

| Loss function | $\mathcal{F}$[%]<br>Left | <br>Right | $\mathcal{L}^{\mathrm{lv}}$<br>$(\times 10^4)$ | $\mathcal{L}^{\mathrm{wd}}$<br>$(\times 10^4)$ | $\mathcal{L}^{\mathrm{ds}}$<br>$(\times 10^4)$ |
|---|---|---|---|---|---|
| $\mathcal{L}^{\mathrm{nt}}$ | 25.6 | 56.1 | 20 | 26 | 0.78 |
| $\mathcal{L}^{\mathrm{nt}}+\beta^{\mathrm{lv}}\mathcal{L}^{\mathrm{lv}}$ | 26.6 | 59.3 | **8** | 15 | 0.75 |
| $\mathcal{L}^{\mathrm{nt}}+\beta^{\mathrm{wd}}\mathcal{L}^{\mathrm{wd}}$ | 26.4 | 58.5 | 10 | 19 | 0.80 |
| $\mathcal{L}^{\mathrm{nt}}+\beta^{\mathrm{ds}}\mathcal{L}^{\mathrm{ds}}$ | 27.2 | 56.4 | 33 | 42 | **0.54** |
| $\mathcal{L}$ | **27.8** | **59.7** | 10 | **13** | 0.67 |

appropriate piano score by directly reducing an original band score. In contrast, 94% of the piano notes were derived from the augmented band scores. This supports our assumption that a reasonable piano score can be obtained by re-using, shifting notes and deleting unnecessary notes from an augmented band score.

The tatum-level onset matching rates (denoted by $\mathcal{F}$) between the estimated and ground-truth piano scores (higher is better) and the statistic-level losses $\mathcal{L}^{\mathrm{lv}}$, $\mathcal{L}^{\mathrm{wd}}$, and $\mathcal{L}^{\mathrm{ds}}$ (lower is better) are shown in Table 1. The best matching rate was achieved when the total loss $\mathcal{L}$ consisting of all three statistic-level losses was minimized. We confirmed the effectiveness of each statistic-level loss in improving $\mathcal{F}$ and reducing $\mathcal{L}^{\mathrm{lv}}$, $\mathcal{L}^{\mathrm{wd}}$, and $\mathcal{L}^{\mathrm{ds}}$. Although the statistical regularization does not aim to directly make the estimated piano score close to the ground-truth score, it was experimentally proven to improve $\mathcal{F}$. This is considered a noticeable improvement.

The distributions of the note statistics obtained from the ground-truth and estimated piano scores are shown in Fig. 5. This indicates that the difficulty level is well characterized by these distributions and supports our difficulty-dependent statistic-level regularization. The distributions for the estimated right-hand parts were close to the ground-truth, whereas those for the estimated left-hand parts were tailed toward larger values than the ground-truth, as shown below.

Examples of piano arrangements are shown in Fig. 6 (other examples are at our webpage[1]). The use of the statistic-level losses made the difference between the estimated elementary and advanced piano scores clearer. The right-hand part had a good correspondence with the melody part, whereas the left-hand part with the advanced level tended to include too many notes and the polyphony width of concurrent notes was often over one octave.

## 5. CONCLUSION

This paper described a difficulty-aware band-to-piano score arrangement method based on the regularized training of a mask estimation U-Net that selects notes from an augmented band score. This score reduction strategy is supported by the statistical investigation of how the left- and right-hand parts of piano scores are derived from the melody and accompaniment parts of band scores. To address the non-uniqueness of the "ground-truth" arrangement, the U-Net is trained with note- and statistic-level criteria such that the polyphony levels, polyphony widths, and note densities of the estimated piano scores are distributed according to a specified difficulty level. Experimental results showed the effectiveness of these note statistics in yielding locally and globally coherent piano scores.

Several open problems remain as future work. To prevent the left-hand part from including unplayable concurrent pitches wider than the hand size, one could reduce the weights of octave-shifted notes in the augmented band scores. In addition, we plan to investigate how the note statistics depend on arrangers and attempt multi-faceted continuous difficulty control based on tempo and key information. We also plan to conduct a subjective evaluation comparing the proposed and conventional methods.

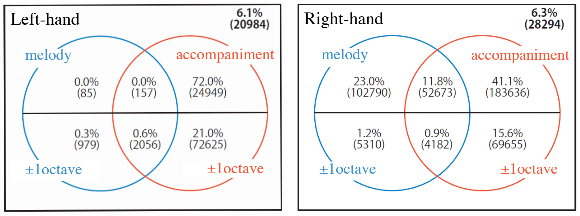

**Fig. 4**. The origins of the left- and right-hand notes of piano scores. The upper area of each blue circle indicates the percentage of piano notes directly derived from melody notes, and the lower area indicates the percentage of piano notes derived from octave-shifted melody notes. The orange circles are defined in the same way for accompaniment notes.

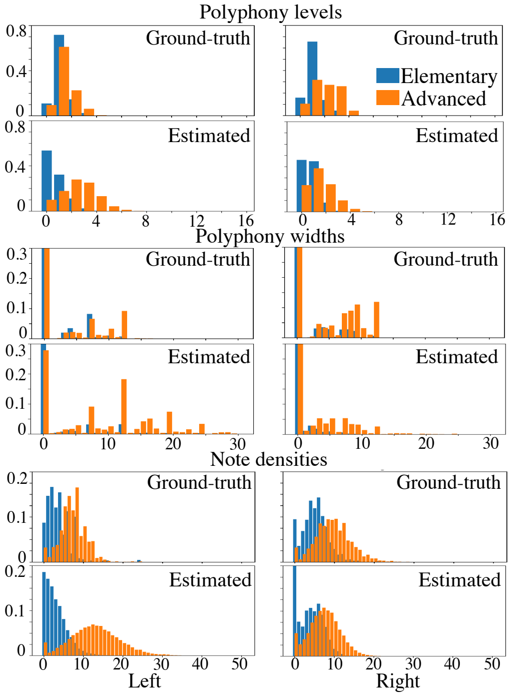

**Fig. 5**. The distributions (histograms) of the note statistics obtained from the ground-truth and estimated piano scores.

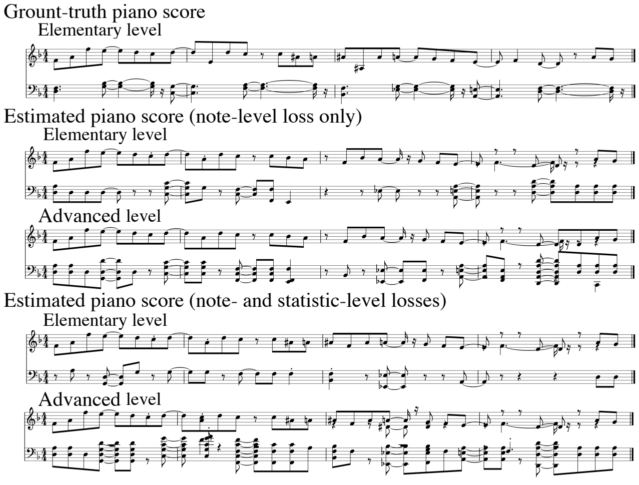
Grount-truth piano score
Elementary level
Estimated piano score (note-level loss only)
Elementary level
Advanced level
Estimated piano score (note- and statistic-level losses)
Elementary level
Advanced level

**Fig. 6**. Examples of automatic band-to-piano score arrangement.

[1] https://teraomoyu.github.io/statistical-piano-arrangement.github.io/

## 6. REFERENCES

[1] Shih-Chuan Chiu, Man-Kwan Shan, and Jiun-Long Huang, "Automatic system for the arrangement of piano reductions," in *Proc. International Symposium on Multimedia*, 2009, pp. 459–464.

[2] Sho Onuma and Masatoshi Hamanaka, "Piano arrangement system based on composers' arrangement processes," in *Proc. International Computer Music Conference*, 2010, pp. 191–194.

[3] Jiun-Long Huang, Shih-Chuan Chiu, and Man-Kwan Shan, "Towards an automatic music arrangement framework using score reduction," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 8, no. 1, pp. 1–23, 2012.

[4] Hirofumi Takamori, Haruki Sato, Takayuki Nakatsuka, and Shigeo Morishima, "Automatic arranging musical score for piano using important musical elements," in *Proc. Sound and Music Computing Conference*, 2017, pp. 35–41.

[5] Eita Nakamura and Kazuyoshi Yoshii, "Statistical piano reduction controlling performance difficulty," *APSIPA Transactions on Signal and Information Processing*, vol. 7, no. e13, pp. 1–12, 2018.

[6] Daniel Tuohy and Walter Potter, "A genetic algorithm for the automatic generation of playable guitar tablature," in *Proc. International Computer Music Conference*, 2005, pp. 499–502.

[7] Gen Hori, Yuma Yoshinaga, Satoru Fukayama, Hirokazu Kameoka, and Shigeki Sagayama, "Automatic arrangement for guitars using hidden Markov model," in *Proc. Sound and Music Computing Conference*, 2012, pp. 450–456.

[8] Gen Hori, Hirokazu Kameoka, and Shigeki Sagayama, "Input-output HMM applied to automatic arrangement for guitars," *Journal of Information Processing*, vol. 21, no. 2, pp. 264–271, 2013.

[9] Hiroshi Maekawa, Norio Emura, Masanobu Miura, and Masuzo Yanagida, "On machine arrangement for smaller wind-orchestras based on scores for standard wind-orchestras," in *Proc. International Conference on Music Perception and Cognition*, 2006, pp. 268–273.

[10] Léopold Crestel and Philippe Esling, "Live orchestral piano, a system for real-time orchestral music generation," in *Proc. Sound and Music Computing Conference*, 2017, pp. 434–442.

[11] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Xianbin Gu, and Gus Xia, "Pop909: A pop-song dataset for music arrangement generation," in *Proc. International Society for Music Information Retrieval Conference*, 2020, pp. 38–45.

[12] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang, "MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proc. AAAI Conference on Artificial Intelligence*, 2018, pp. 34–41.

[13] Yun-Ning Hung, I-Tung Chiang, Yi-An Chen, and Yi-Hsuan Yang, "Musical composition style transfer via disentangled timbre representations," in *Proc. International Joint Conference on Artificial Intelligence*, 2019, pp. 4697–4703.

[14] Eita Nakamura and Shigeki Sagayama, "Automatic piano reduction from ensemble scores based on merged-output hidden Markov model," in *Proc. International Computer Music Conference*, 2015, pp. 298–305.

[15] Eita Nakamura, Nobutaka Ono, and Shigeki Sagayama, "Merged-output HMM for piano fingering of both hands," in *Proc. International Society for Music Information Retrieval Conference*, 2014, pp. 531–536.

[16] Yuichiro Yonebayashi, Hirokazu Kameoka, and Shigeki Sagayama, "Automatic decision of piano fingering based on a hidden Markov models," in *Proc. International Joint Conference on Artificial Intelligence*, 2007, pp. 2915–2921.

[17] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck, "Music transformer: Generating music with long-term structure," in *Proc. International Conference on Learning Representations*, 2019.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[19] Eric Jang, Shixiang Gu, and Ben Poole, "Categorical reparameterization with Gumbel-Softmax," in *Proc. International Conference on Learning Representations*, 2017.

[20] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations*, 2015.