

Prenošenje stila glazbe korištenjem difuzijskih modela

PROJEKT IZ KOLEGIJA DUBOKO UČENJE 2

FILIP PANKRETIĆ, FILIP PERKOVIĆ, FRAN VUČKOVIĆ, LUKA GLAVINIĆ, VELIMIR
KOVAČIĆ, DOMINIK JAMBROVIĆ

FER 2024./2025.

Sadržaj

1. Uvod
2. Skup podataka Pixabay
3. Arhitektura, učenje i prijenos stila
4. Mjere dobrote
5. Eksperimentalni rezultati
6. Usporedba s prijašnjim radovima
7. Zaključak

Opis problema i motivacija

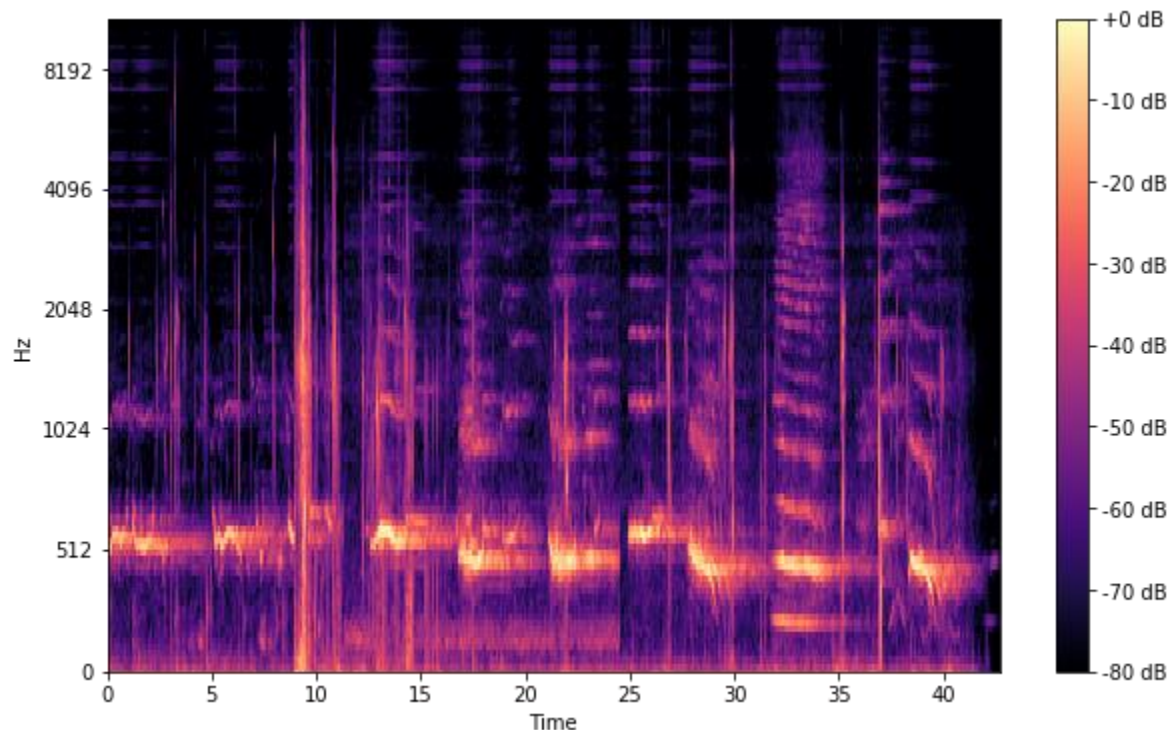
- Ulaz:
 - Zvučni zapis(i) stila
 - Zvučni zapis sadržaja
- Izlaz:
 - Stilizirani zvučni zapis
- Mogućnost automatiziranog generiranja stiliziranih pjesama bez potrebe za detaljnim tekstualnim opisima stila

Postojeća rješenja

- Riffusion, MUSICGEN, SS VQ-VAE, ...
- Većina dosadašnjih rješenja može izvrsno provesti prijenos stila, ali samo za stilove koje je model vidio tijekom učenja
- Dodatno, često neuspješan prijenos stilova zvukova prirode ili umjetno generiranih zvuka
- Modeli često kao ulaz očekuju detaljan tekstualni opis stila

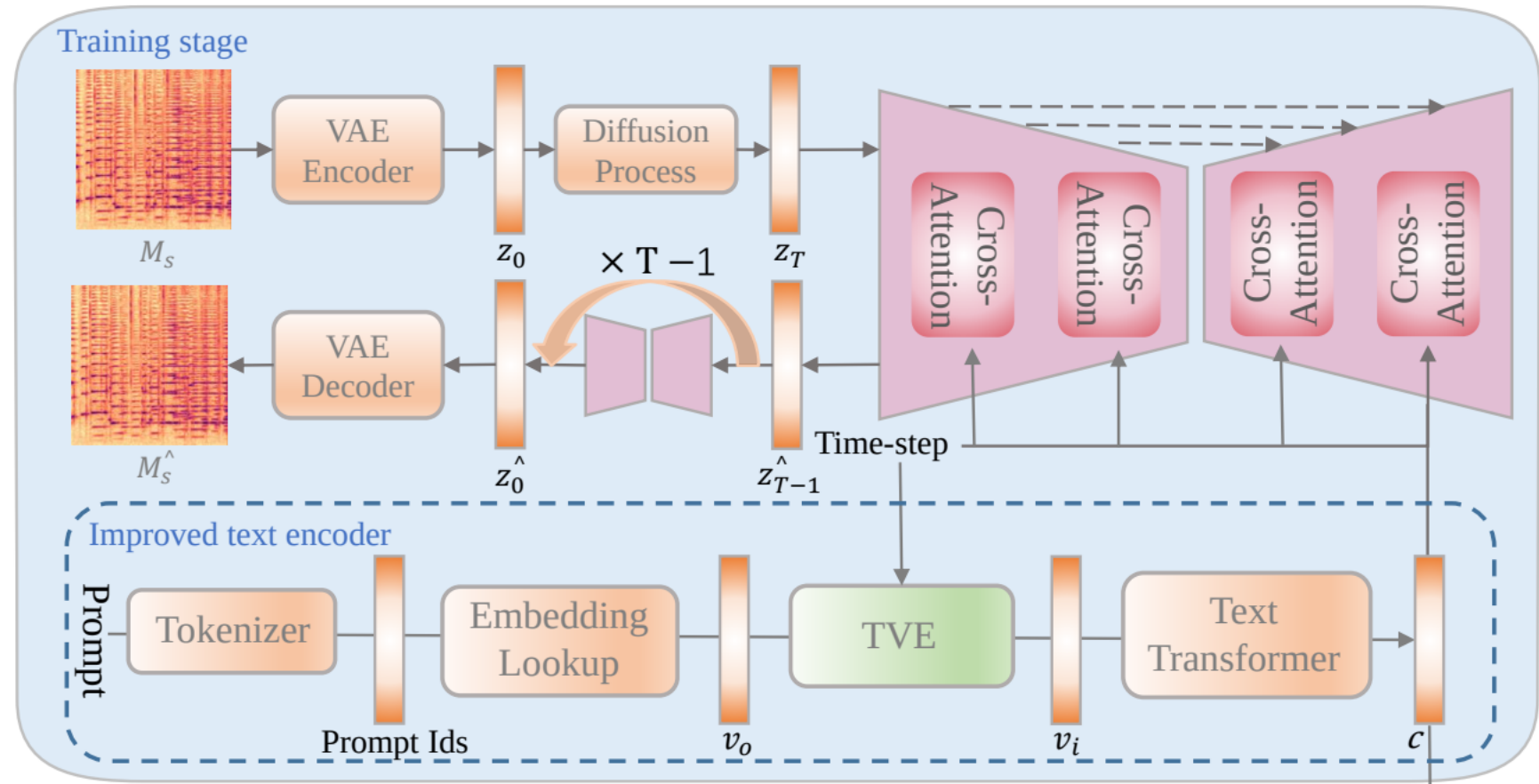
Skup podataka Pixabay

- 253 zvučna zapisa .wav formata (pretvorba u mel-spektrogram)
- Podjela:
 - Sadržaj:
 - 179 zapisa
 - 13 kategorija
 - Stil
 - 74 zapisa
 - 18 kategorija



Arhitektura

- Okosnica Riffusion
- Koder teksta iz CLIP-a
- Modul TVE



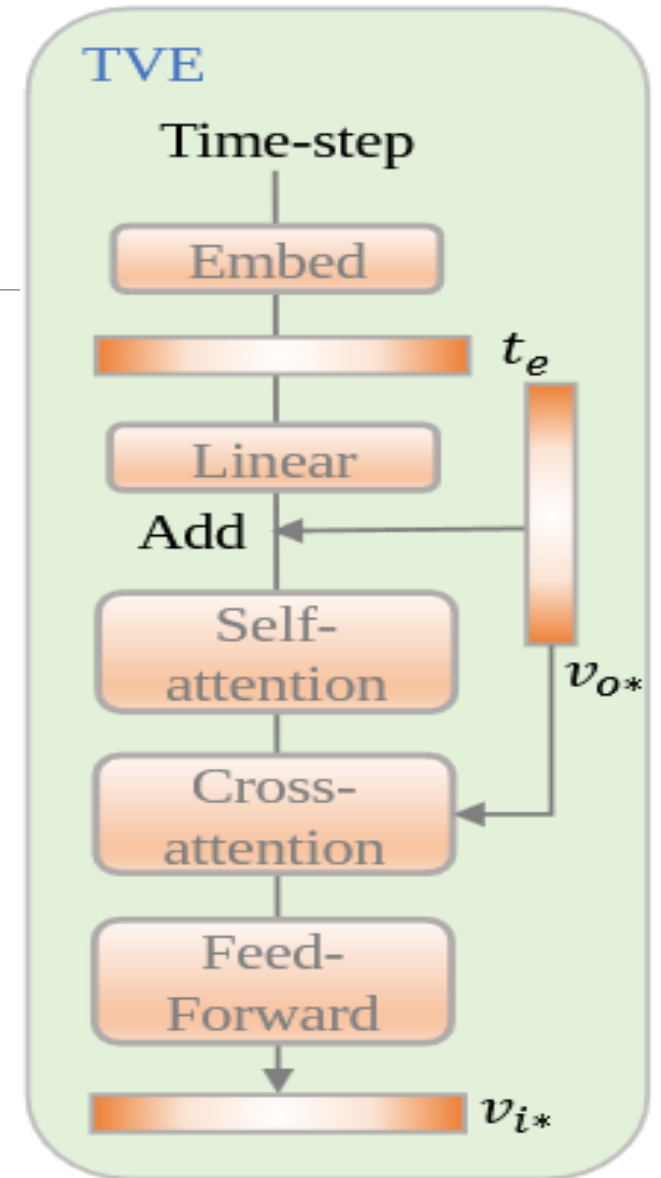
Arhitektura, modul TVE

- Cilj: asocirati vremenski korak t i ugrađivanje znaka "*" sa stilom tijekom učenja

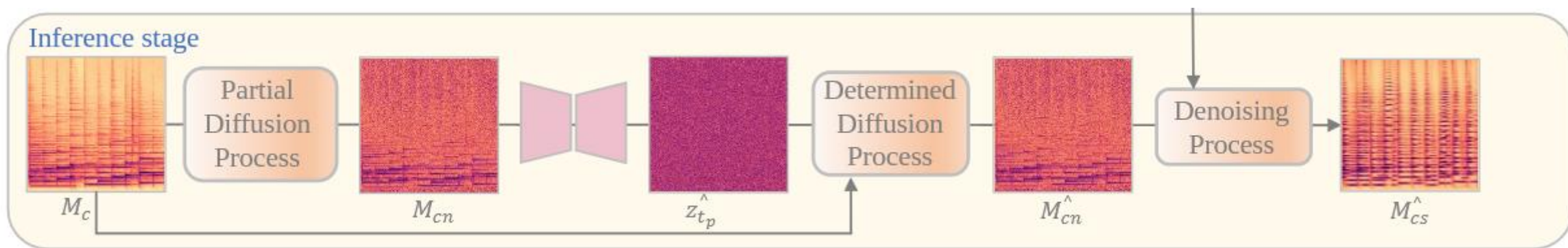
- Uče se isključivo parametri ovog modula

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{z,y,\epsilon_t,t} [\|\epsilon_t - \epsilon_{\phi}(z_t, t, e_{\theta}(y, t))\|_2^2]$$

- Nakon učenja, modul vodi proces prijenosa stila



Prijenos stila



- Hiperparametri: broj koraka prijenosa, scale, strength

$$\hat{\epsilon}_t = \hat{\epsilon}_{t,uncod} + scale \cdot (\hat{\epsilon}_{t,text} - \hat{\epsilon}_{t,uncod})$$

Mjere dobrote

- Mjera CLAP

$$\text{CLAPscore}(x, y) = \text{ReLU} \left(\frac{E(x) \cdot E(y)}{\|E(x)\| \cdot \|E(y)\|} \right)$$

- Očuvanje sadržaja

$$CP(x_0, \hat{x}_0) = \text{CLAPscore}(x_0, \hat{x}_0)$$

- Podudaranje stila

$$SF(\hat{x}_0, Y) = \text{ReLU} \left(\frac{E(\hat{x}_0) \cdot \frac{1}{N} \sum_{i=1}^N E(y_i)}{\|E(\hat{x}_0)\| \cdot \|\frac{1}{N} \sum_{i=1}^N E(y_i)\|} \right)$$

Eksperimenti

Eksperimentalne postavke

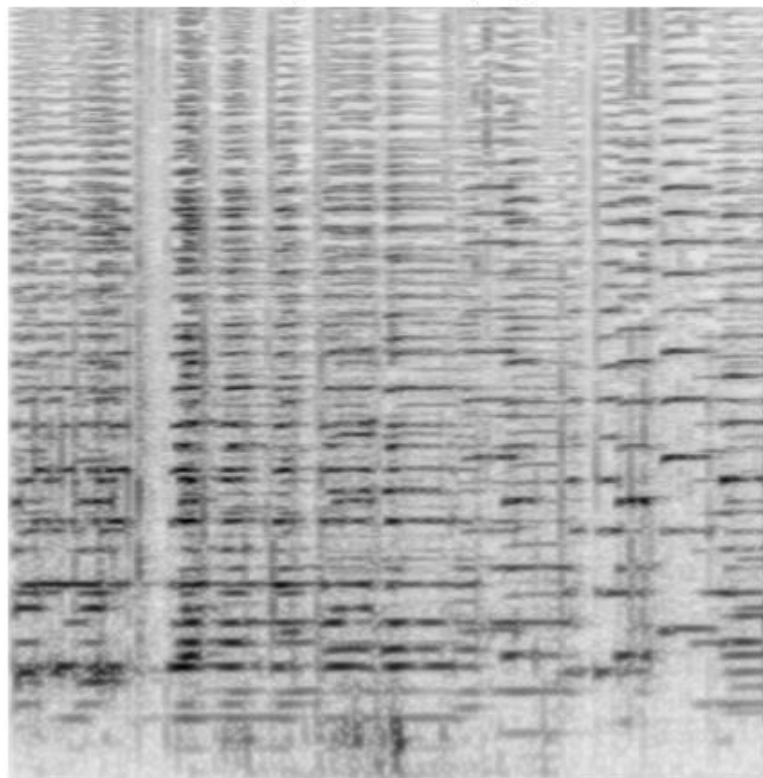
- 1 stil = 1 model, 3 odabrana stila: *bird*, *accordion*, *chime*
- Optimizator Adam:
 - Početni LR: 0.0001
 - β_1 : 0.9, β_2 : 0.999
- 3000 epoha, mini-grupe veličine 1

Eksperimentalni rezultati

Stil	Scale	Strength	CP	SF
<i>Accordion</i>	4.5	0.6	0.40	0.49
<i>Bird</i>	3.5	0.45	0.35	0.40
<i>Chime</i>	3.5	0.45	0.49	0.41
Prosjek	-	-	0.41	0.43

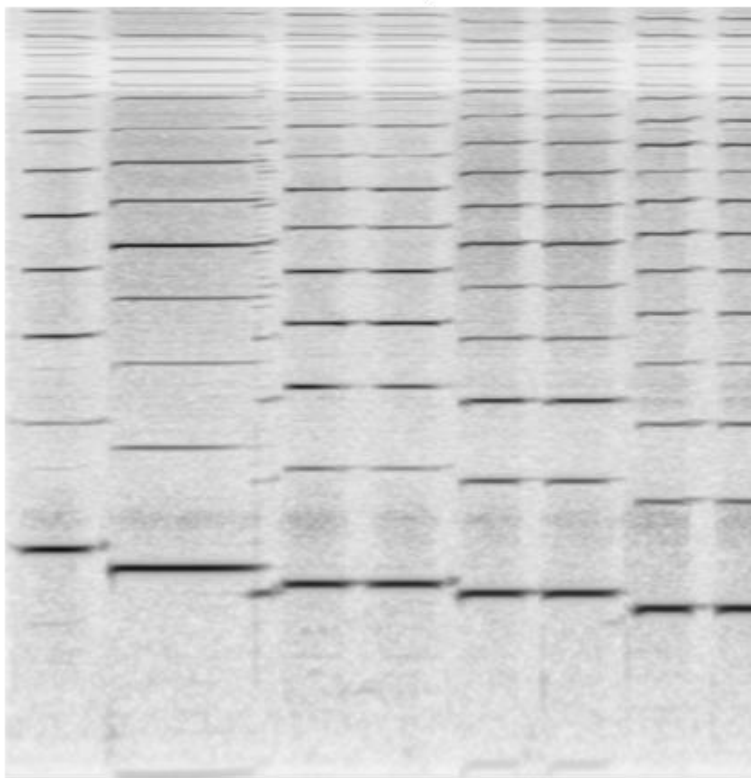
Eksperimentalni rezultati

Stil (accordion7.png)

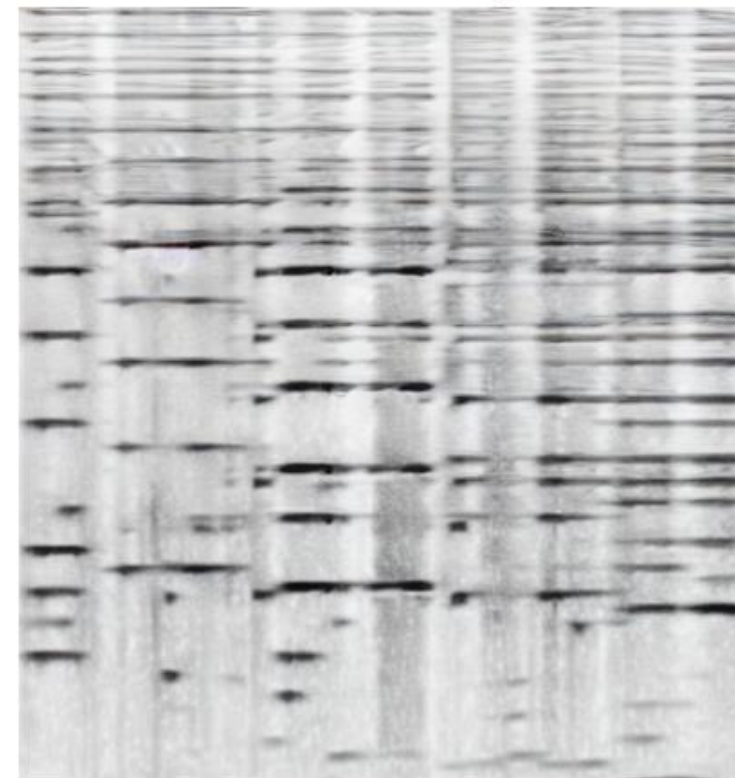


Najbolji prijenos stila (0.79), twinkle2.png

Sadržaj



Stilizirani sadržaj



Eksperimentalni rezultati - demonstracija

- Originalni zvučni zapisi



- Stilizirani zvučni zapisi

Accordion



Bird



Heartbeat



Usporedba s prijašnjim radovima

Model	CP	SF
R+TI [11] [13]	0.35	0.27
SS VQ-VAE [14]	0.24	0.28
MUSICGEN [15]	0.28	0.24
Originalni model [3]	0.46	0.28
Naš model	0.41	0.43

Zaključak

- Pristup temeljen na korištenju latentnog difuzijskog modela uz koder teksta iz CLIP-a i modul TVE predstavlja obećavajuć smjer rada u području prijenosa stila glazbe
- Model ne očekuje tekstualni opis stila, već je dovoljno prikupiti nekoliko kratkih zvučnih zapisa
- Učenje modela moguće provoditi u razumnom vremenu na osobnom računalu

Hvala na pozornosti!
