iSTFTNet: Fast and Lightweight Mel-Spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform

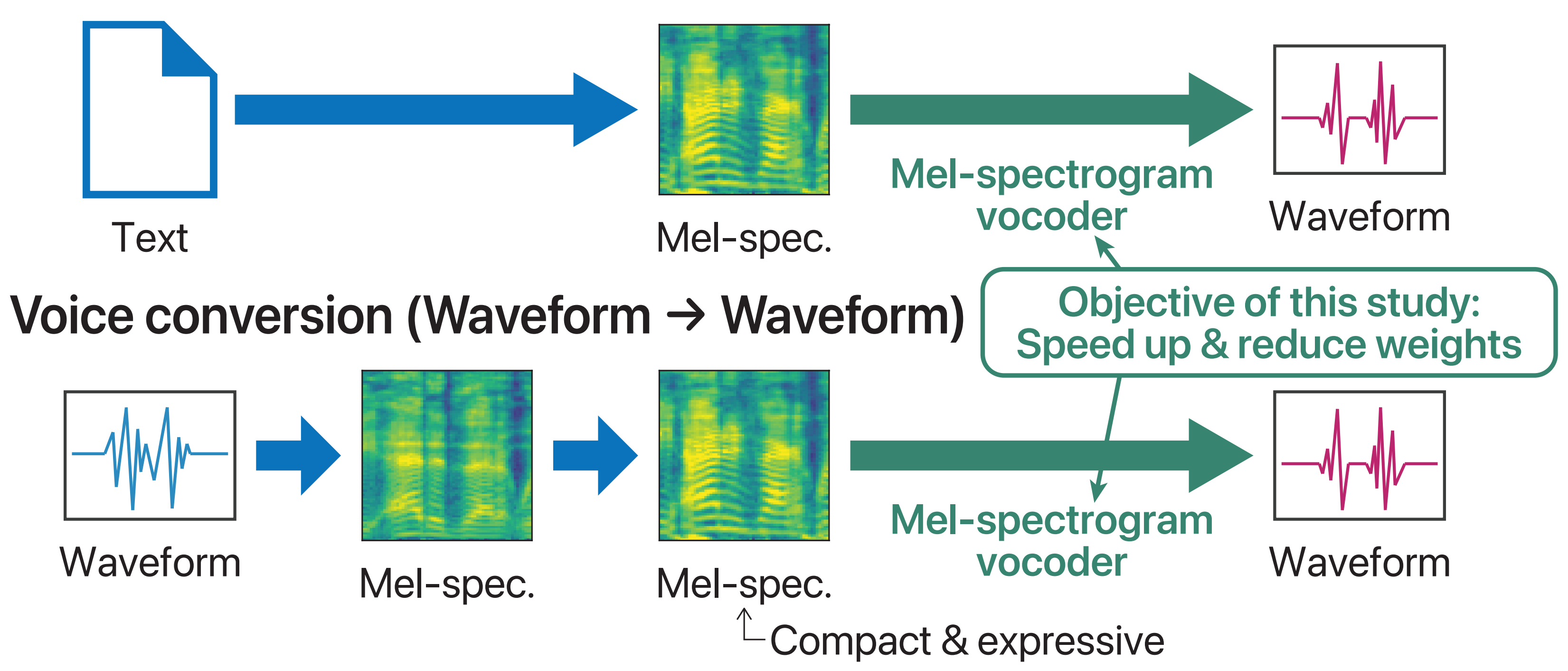
Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, Shogo Seki
NTT Communication Science Laboratories, NTT Corporation, Japan

SPE-7.5
IEEE ICASSP 2022

Audio samples
https://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/istftnet/

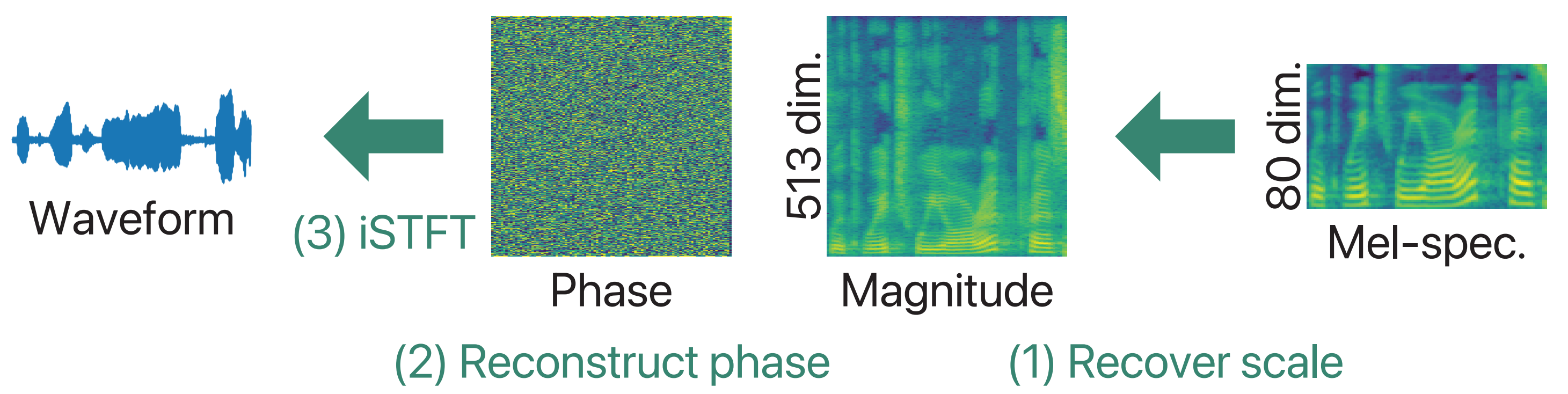
❶ Background

Increased demand for efficient mel-spectrogram vocoder



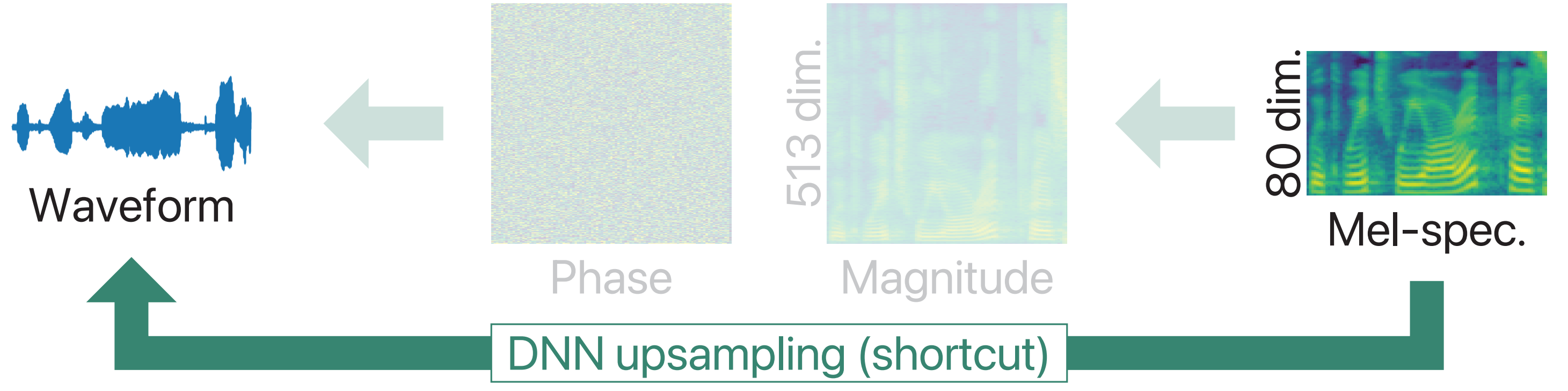
Typical mel-spectrogram vocoders

Signal processing-based solution



Pros: Exploits time-frequency structure explicitly
Cons: Requires redundant estimation (reconstruction of high-dim. spec.)

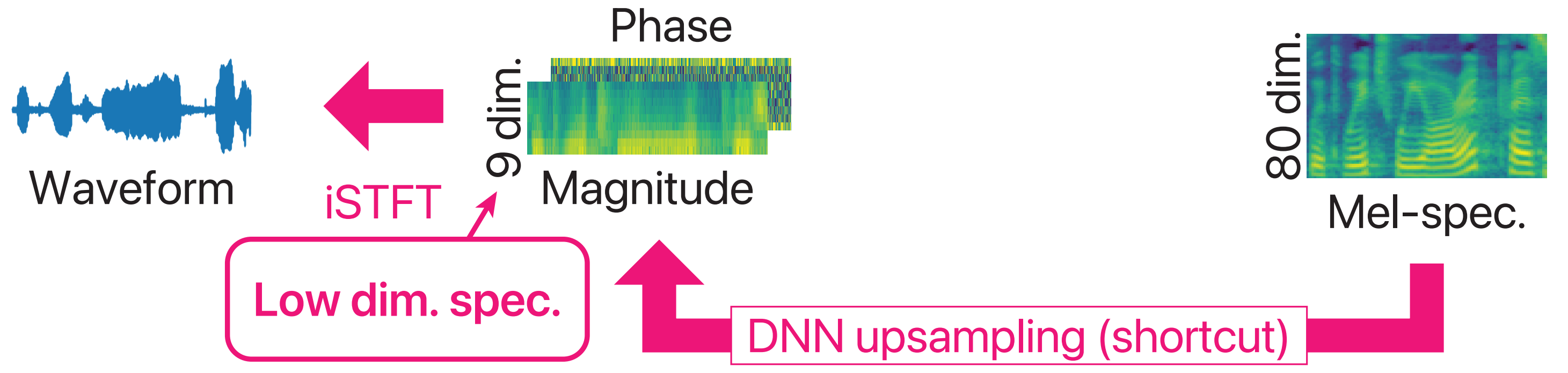
DNN-based shortcut solution



Pros: Does not require redundant estimation (reconstruction of high-dim. spec.)
Cons: Cannot exploit time-frequency structure explicitly

❷ Key idea: Hybrid approach

Utilization of both strengths



Pros: Avoids redundant estimation using **DNN upsampling**
Pros: Exploits time-frequency structure explicitly using **iSTFT**

❸ Theoretical Background

Time-frequency trade-off

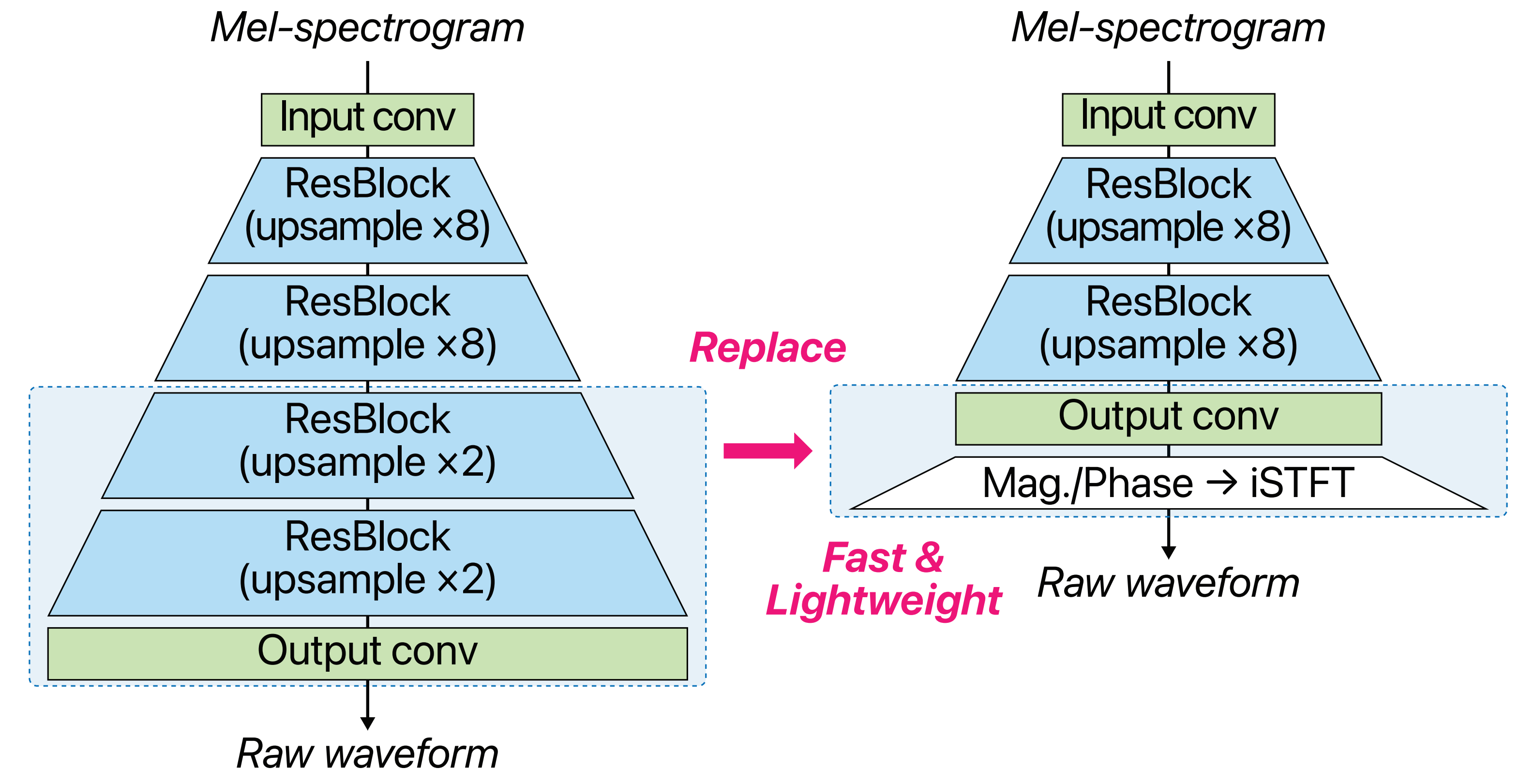
$$f_1 \cdot 1 = f_s \cdot s = \text{constant}$$

FFT size Time scale

We can simplify frequency structure by increasing time scale

❹ Proposal: *iSTFTNet*

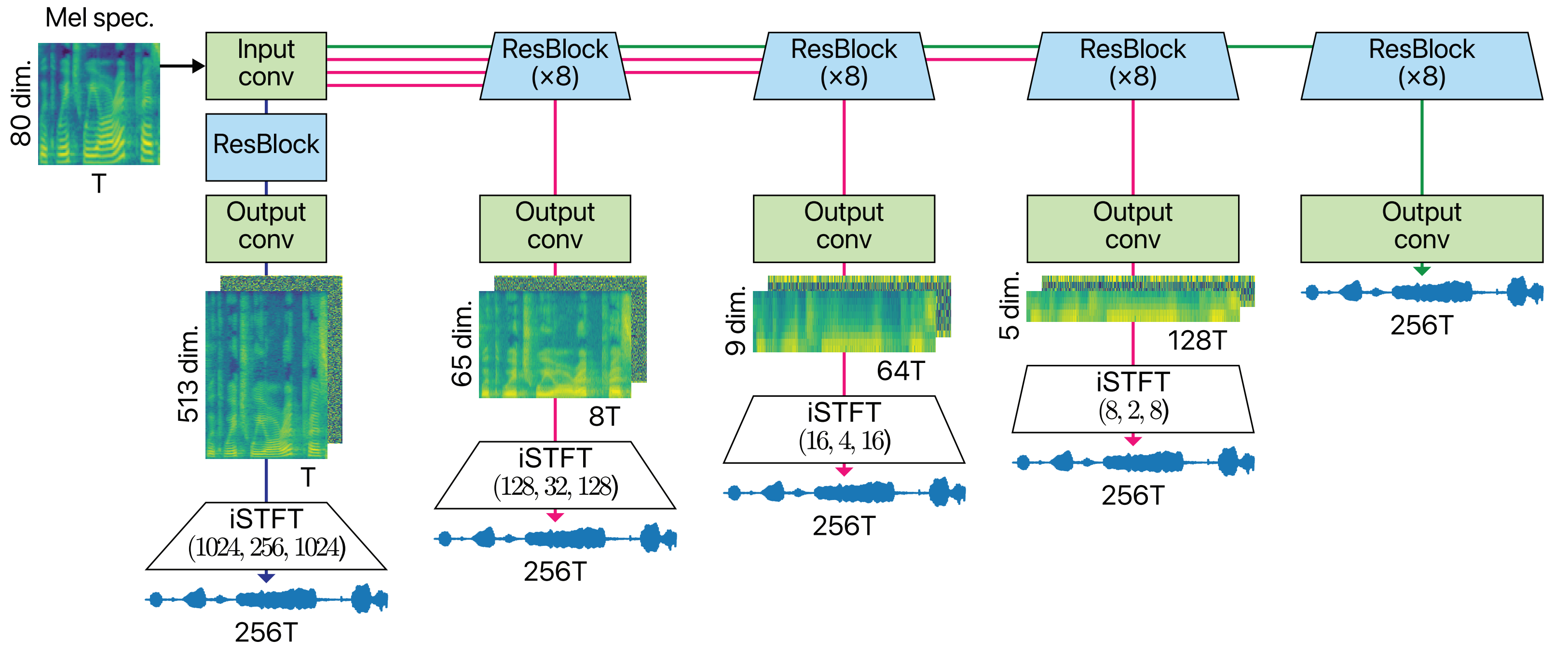
Hybrid of DNN upsampling & iSTFT signal processing



(a) Standard mel-spectrogram vocoder (b) *iSTFTNet* (proposed)

1. **Simplifies frequency structure** using **DNN upsampling**
2. **Exploits time-frequency structure** explicitly using **iSTFT**

❺ Architectures of *iSTFTNets*



(a) C1I (No upsampling) (b) C8I (c) C8C8I (d) C8C8C2I (e) C8C8C2C2 (Previous)

iSTFTNets

Fast Lightweight Low quality? ⟷ Slow Heavyweight High quality?

We examined effect on quality empirically

❻ Experiments

Experiment settings

Dataset: LJSpeech [Ito&Johnson+17]
- **Speaker:** English female speaker
- **Audio clips:** 13,100 (24 h) (training: 12,500, validation: 250, evaluation: 250)
- **Sampling rate:** 22.05 kHz
- **Audio features:** Log-mel spectrogram (FFT: 1024, hop: 256, window: 1024)

Evaluation metrics:
- **MOS↑:** Mean opinion score on naturalness (from 1 (bad) to 5 (excellent))
- **cFW2VD↓:** Distance between real & generative distributions in wav2vec 2.0
- **Speed↑:** Relative speed compared to real time on GPU/CPU
- **#Param↓:** Number of parameters

Comparison models
- **HiFi-GANs** [Kong+2020]: **V1** (high-quality), **V2** (lightweight), **V3** (fast)
- **Multiband (MB)-MelGAN** [Yang+2021], **Parallel WaveGAN (PWG)** [Yamamoto+2020]

Results (Synthesis from ground-truth mel-spectrogram)

Q1. How many blocks should be retained?

# Retained layers	Model	MOS↑	cFW2VD↓	Speed on GPU↑	Speed on CPU↑	# Param (M)↓
	Ground truth	4.46 ±0.14	–	–	–	–
4	V1 (original)	4.22 ±0.17	0.020	×143.59 (100)	×1.34 (100)	13.94 (100)
3	V1-C8C8C2I	4.22 ±0.17	0.018	×179.42 (125)	×1.63 (122)	13.80 (99)
2	V1-C8C8I	4.26 ±0.17	0.020	×245.68 (171)	×2.33 (174)	13.26 (95)
1	V1-C8I	3.32 ±0.22	0.073	×609.43 (424)	×7.57 (565)	10.89 (78)
4	V2 (original)	3.91 ±0.17	0.046	×624.47 (100)	×10.39 (100)	0.93 (100)
3	V2-C8C8C2I	3.98 ±0.17	0.038	×732.96 (117)	×13.34 (128)	0.92 (99)
2	V2-C8C8I	3.95 ±0.16	0.042	×1025.46 (164)	×20.37 (196)	0.89 (96)
1	V2-C8I	3.21 ±0.20	0.096	×1720.91 (276)	×68.05 (655)	0.78 (84)
3	V3 (original)	3.78 ±0.16	0.052	×933.06 (100)	×10.40 (100)	1.46 (100)
2	V3-C8C8I	3.41 ±0.19	0.055	×1517.70 (163)	×21.48 (206)	1.42 (97)
1	V3-C8I	2.89 ±0.17	0.156	×2481.87 (266)	×66.83 (642)	1.28 (87)

We can make the models *faster* and *more lightweight* with *reasonable quality* when *3 or 2 blocks are retained*

Q2. Necessity of combining DNN upsampling & iSTFT

Upsampling	Model	MOS↑	cFW2VD↓	Speed on GPU↑	Speed on CPU↑	# Param (M)↓
2	V1-C8C8I	4.26 ±0.17	0.020	×245.68 (171)	×2.33 (174)	13.26 (95)
1	V1-C8C1I	3.82 ±0.17	0.033	×326.39 (227)	×3.97 (296)	19.15 (137)
2	V2-C8C8I	3.95 ±0.16	0.042	×1025.46 (164)	×20.37 (196)	0.89 (96)
1	V2-C8C1I	3.44 ±0.20	0.071	×1081.37 (173)	×39.14 (377)	1.30 (140)
2	V3-C8C8I	3.41 ±0.19	0.055	×1517.70 (163)	×21.48 (206)	1.42 (97)
1	V3-C8C1I	2.82 ±0.21	0.116	×1925.15 (206)	×41.16 (396)	1.77 (121)

Upsampling is necessary to *preserve quality*

Q3. Comparison with benchmark models

Model	MOS↑	cFW2VD↓	Speed on GPU↑	Speed on CPU↑	# Param (M)↓
V2-C8C8I	3.95 ±0.16	0.042	×1025.46	×20.37	0.89
MB-MelGAN	3.54 ±0.21	0.078	×1070.95	×17.95	2.54
PWG	3.47 ±0.21	0.066	×79.71	×0.70	1.35

Quality & Size: *iSTFTNet* is *best*   Speed: *iSTFTNet* = MB-MelGAN

Application to text-to-speech synthesis

Model	MOS↑	cFW2VD↓
Ground truth	4.32 ±0.10	–
Conformer-FS2 + V1	4.09 ±0.12	0.216
Conformer-FS2 + V1-C8C8I	4.25 ±0.11	0.214
Conformer-FS2 [Guo+2021]	3.66 ±0.15	0.242

- *iSTFTNet* is *better* than or *comparable* with **baselines**
- *iSTFTNet* is *comparable* with **ground truth**