Sony AI

Supervised Contrastive Learning from Weakly-labeled Audio Segments for Musical Version Matching

Joan Serrà[1], **R. Oguz Araz**[2], **Dmitry Bogdanov**[2], **& Yuki Mitsufuji**[1,3]

[1] Sony AI
[2] Music Technology Group, UPF
[3] Sony Group Corporation

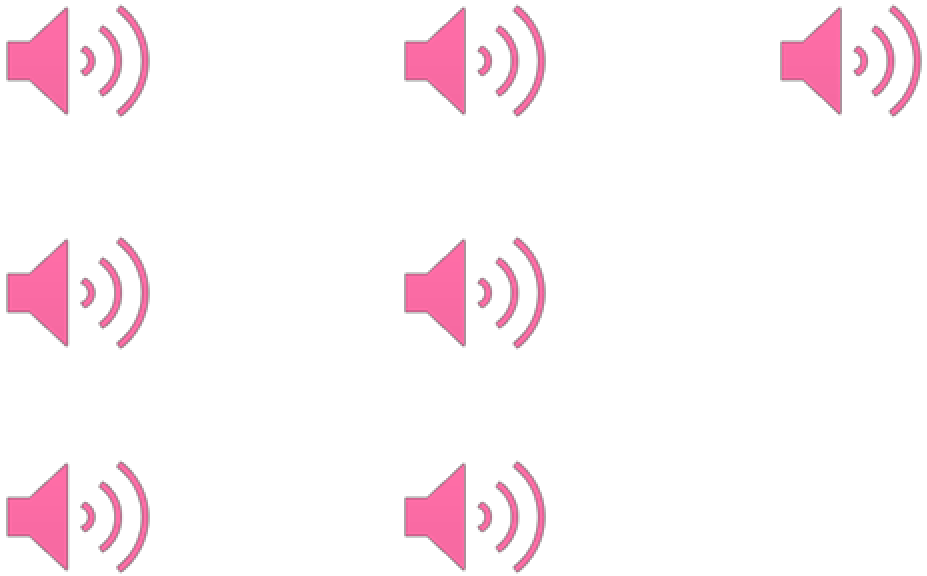
October 2025 – Conversational AI Reading Group

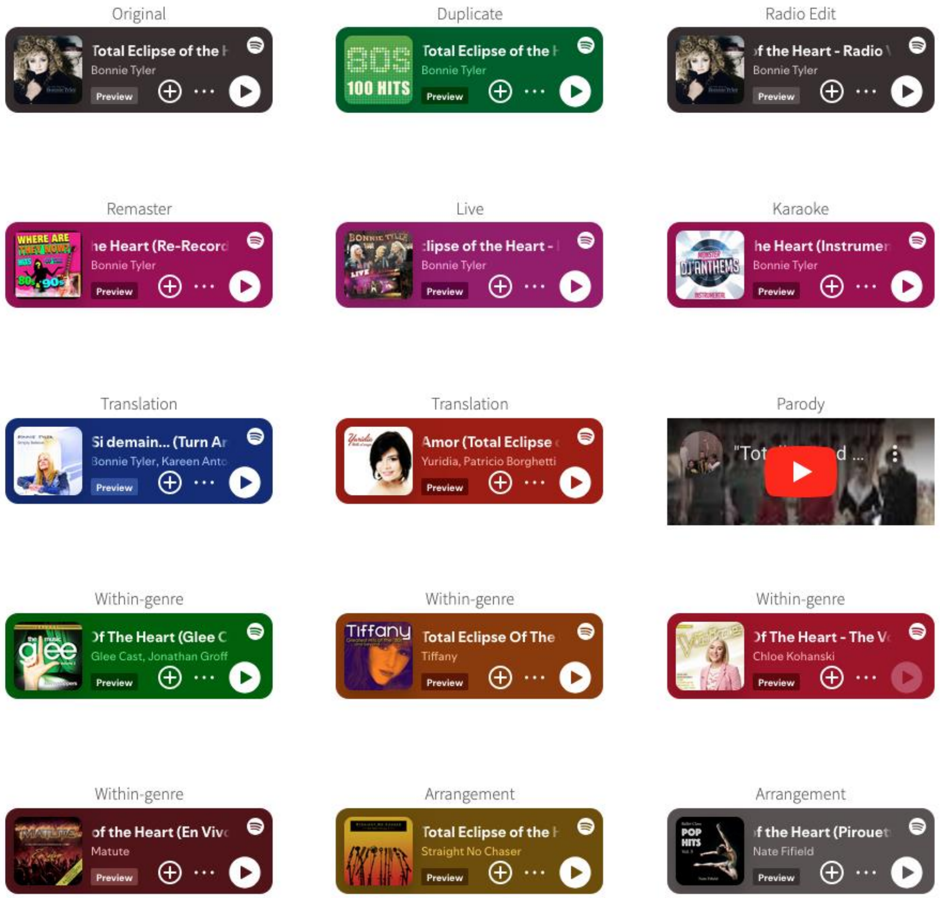
Musical Versions

Different renditions of the same musical piece or passage

Musical Versions

Different renditions of the same musical piece or passage





Further examples: https://secondhandsongs.com/ or https://furkanyesiler.github.io/musical_version_id_spm/

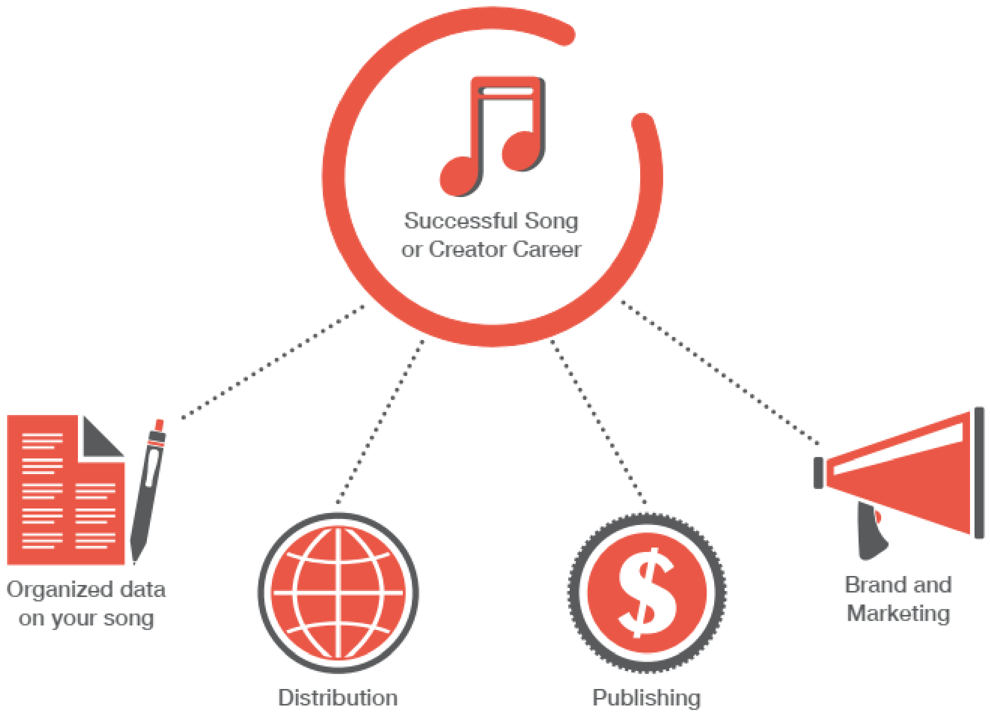
Musical Characteristic	Version Type																	
	Duplicate	Remaster	Radio Edit	Translation	Performance	Demo	Parody	Within-Genre	Karaoke	Live	Standard	Mashup	Acoustic	Medley	Remix	Cross-Genre	Arrangement	Quotation
Melody	0	0	0	0	0	1	0	1	2	1	1	0	1	1	1	2	2	2
Harmony	0	0	0	0	0	1	0	1	0	0	2	0	1	1	2	2	2	3
Tempo	0	0	0	0	2	1	1	1	0	2	1	3	2	2	3	2	2	3
Timing	0	0	0	0	2	1	1	1	0	2	1	3	2	2	2	3	3	3
Structure	0	0	1	0	1	1	1	1	1	2	2	3	2	3	3	2	3	3
Lyrics	0	0	1	3	0	1	3	0	3	1	0	0	0	1	1	1	1	2
Key	0	0	0	0	1	1	1	1	1	1	1	2	2	2	2	3	3	3
Timbre	0	0	0	0	1	1	1	1	2	2	3	2	3	2	3	3	3	3
Noise	0	1	1	1	3	3	2	3	2	3	3	2	3	3	2	3	3	3

Degree of Potential Difference	0	1	2	3
	Likely the Same	May Be Variations	May Be Major Differences	May Be Unrelated

From Yesiler et al. (2021), "Audio-based musical version identification: elements and challenges", IEEE Signal Processing Magazine 38(6): 115-136.

Applications

- Digital rights/copyright management
  - Content monitoring
  - Copyright infringement
- Catalog organization
  - Duplicate/near-duplicate assessment
  - Link related items
- Discovery/creative tool
  - Music recommendation
  - Creative inspiration
  - Preserve/relate cultural heritage



Musical Version Matching

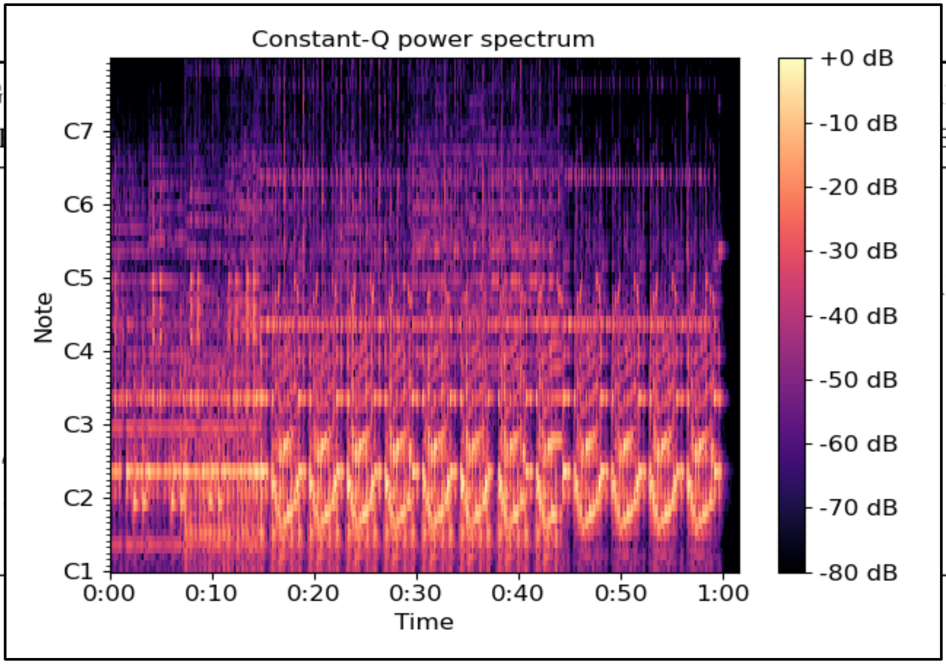
Compute embedding → Store in database → Nearest neighbor retrieval

Musical Version Matching

Compute embedding → Store in database → Nearest neighbor retrieval

Table 1. Comparison of characteristics for a number of existing approaches and the proposed method CLEWS. We exclude multi-feature and/or multi-modal approaches (for example fusing CQT and melody estimations or leveraging audio and lyrics information). For further details and approaches we refer to the survey by Yesiler et al. (2021).

NAME(S)	MAIN REFERENCE	INPUT	ARCH.	SEG LEA	L E
CQTNET	YU ET AL. (2020)	CQT	CONVNET		
DORAS&PEETERS	DORAS & PEETERS (2020)	CQT	CONVNET		
MOVE/RE-MOVE	YESILER ET AL. (2020A)	CREMA	CONVNET		N
PICKINET	O'HANLON ET AL. (2021)	CQT	CONVNET		
LYRACNET	HU ET AL. (2022)	CQT	WIDERESNET		
BYTECOVER1/2	DU ET AL. (2022)	CQT	RESNET		
COVERHUNTER	LIU ET AL. (2023)	CQT	CONFORMER		
BYTECOVER3/3.5	DU ET AL. (2023)	CQT	RESNET		
DVINET/DVINET+	ARAZ ET AL. (2024A)	CQT	CONVNET		



Sony AI

Musical Version Matching

Compute embedding → Store in database → Nearest neighbor retrieval

Table 1. Comparison of characteristics for a number of existing approaches and the proposed method CLEWS. We exclude multi-feature and/or multi-modal approaches (for example fusing CQT and melody estimations or leveraging audio and lyrics information). For further details and approaches we refer to the survey by Yesiler et al. (2021).

Name(s)	Main reference	Input	Arch.	Segment learning	Partial match	Loss / train concept	Retrieval distance
CQTNet	Yu et al. (2020)	CQT	ConvNet	✗	✗	Classif.	Cosine
Doras&Peeters	Doras & Peeters (2020)	CQT	ConvNet	✗	✗	Triplet	Cosine
MOVE/Re-MOVE	Yesiler et al. (2020a)	CREMA	ConvNet	✗	✗	Triplet	Euclidean
PiCKINet	O'Hanlon et al. (2021)	CQT	ConvNet	✗	✗	Classif.+Center	Cosine
LyraCNet	Hu et al. (2022)	CQT	WideResNet	✗	✗	Classif.	Cosine
ByteCover1/2	Du et al. (2022)	CQT	ResNet	✗	✗	Classif.+Triplet	Cosine
CoverHunter	Liu et al. (2023)	CQT	Conformer	∼	✓	Classif.+Focal+Center	Cosine
ByteCover3/3.5	Du et al. (2023)	CQT	ResNet	✓	✗	Classif.+Triplet	Cosine
DVINet/DVINet+	Araz et al. (2024a)	CQT	ConvNet	✗	✗	Triplet	Cosine
CLEWS (proposed)	This paper	CQT	ResNet	✓	✓	Contrastive	Euclidean

Sony AI

Contrastive Learning from Weakly-labeled Segments (CLEWS)

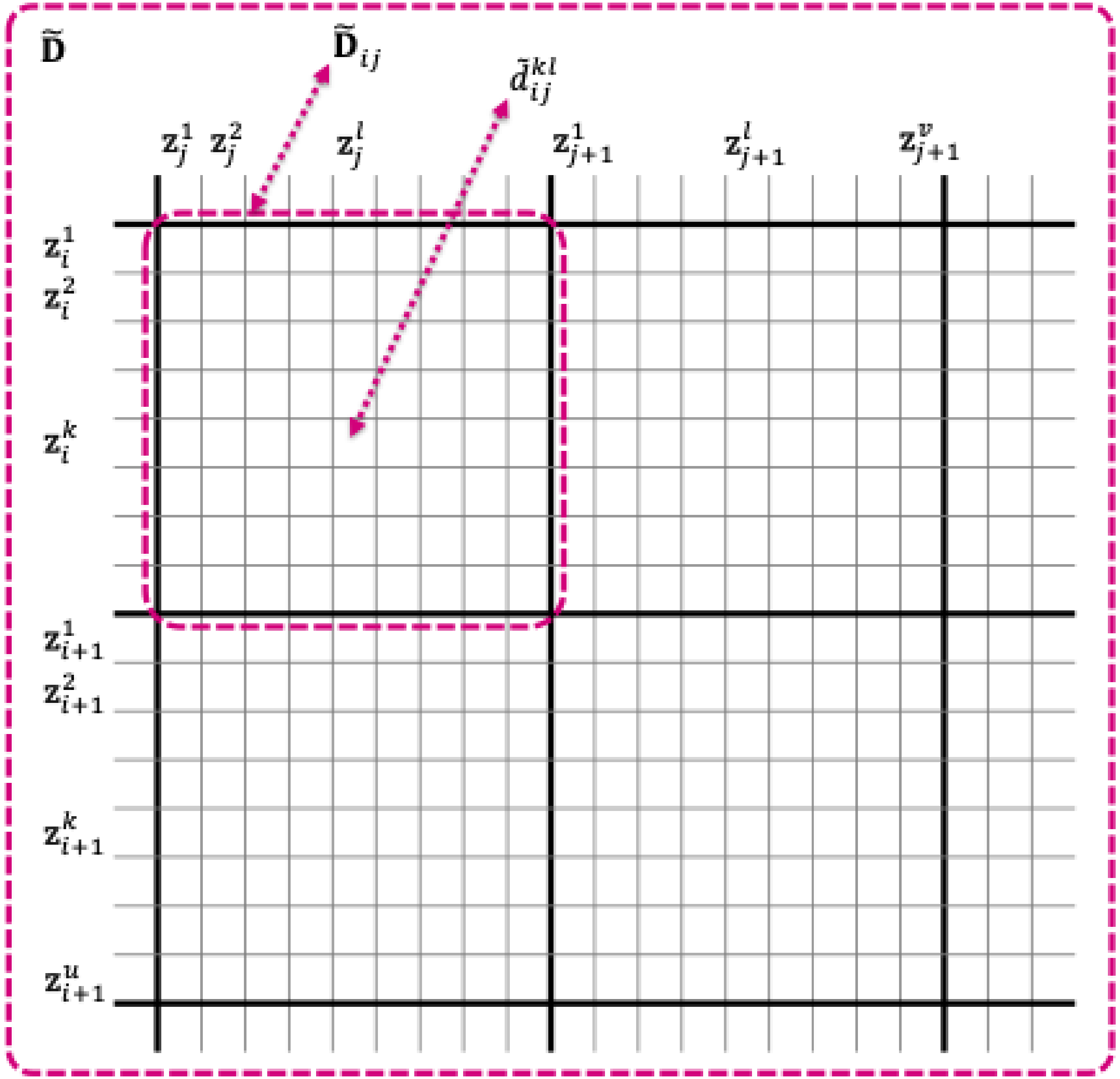
Two main contributions:

- Segment-based learning *and* matching
  - Pairwise segment distance reductions: bpwr-k
  - Different reductions for positive and negative pairs
- Better contrastive loss
  - Evolved from alignment and uniformity (Wang & Isola, 2020)
  - Three new major considerations
    - Decoupling: No overlap between positive and negative pairs
    - Hyper-parameters: Remove/add + Comparable gradient contribution for positive and negative pairs + Soft threshold for “easy” negative pairs
    - Geometric: Space geometry and geodesic distance should match

Wang & Isola (2020), “Understanding contrastive representation learning through alignment and uniformity on the hypersphere”, Proc. of Int. Conf. on Machine Learning (ICML) 119: 9929-9939.

CLEWS: Reductions

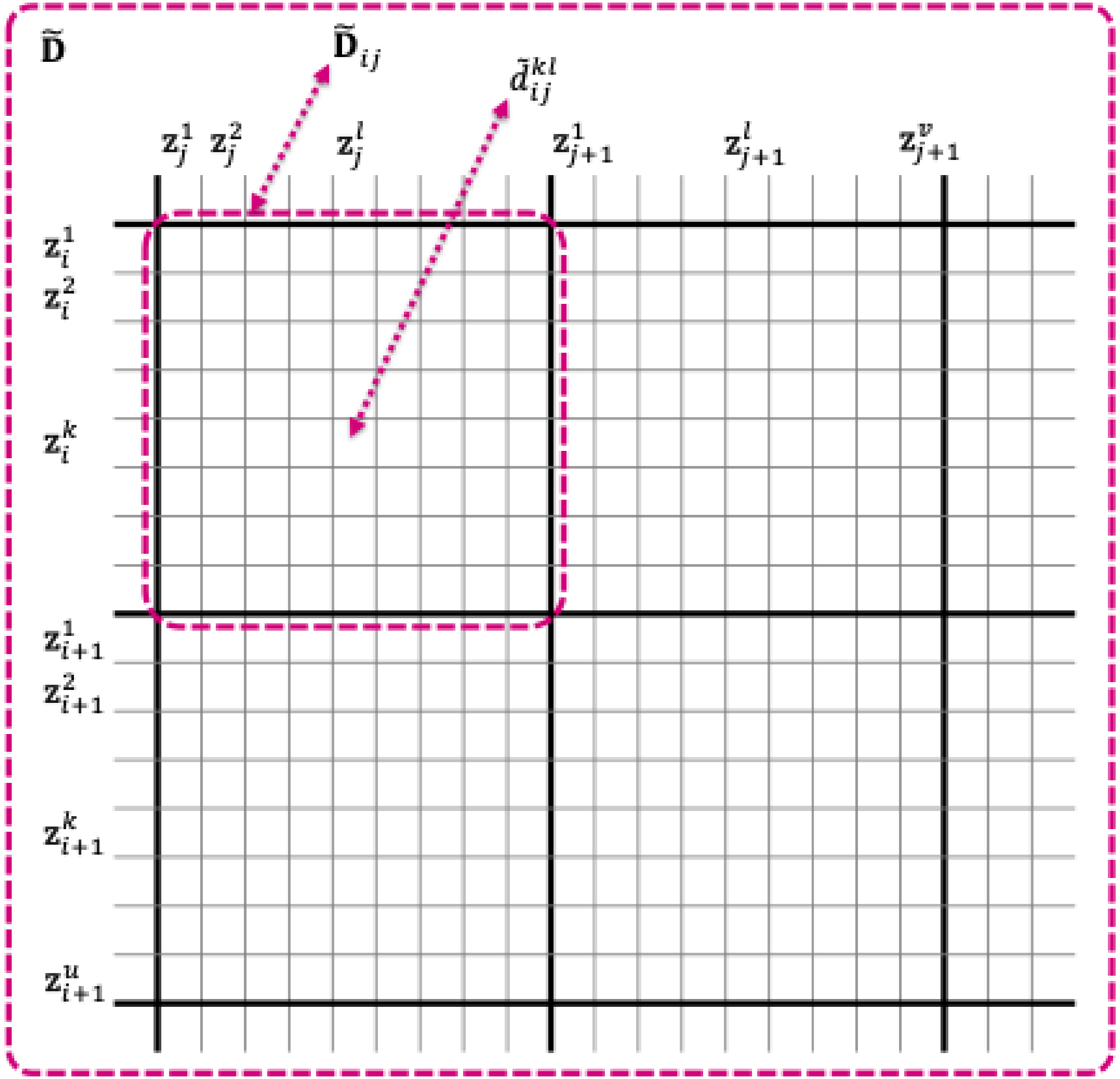
Segment-based learning and matching:



CLEWS: Reductions

Segment-based learning and matching:

- Reduction types: R_{mean}, $R_{\text{top-k}}$, R_{meanmin}, R_{min}

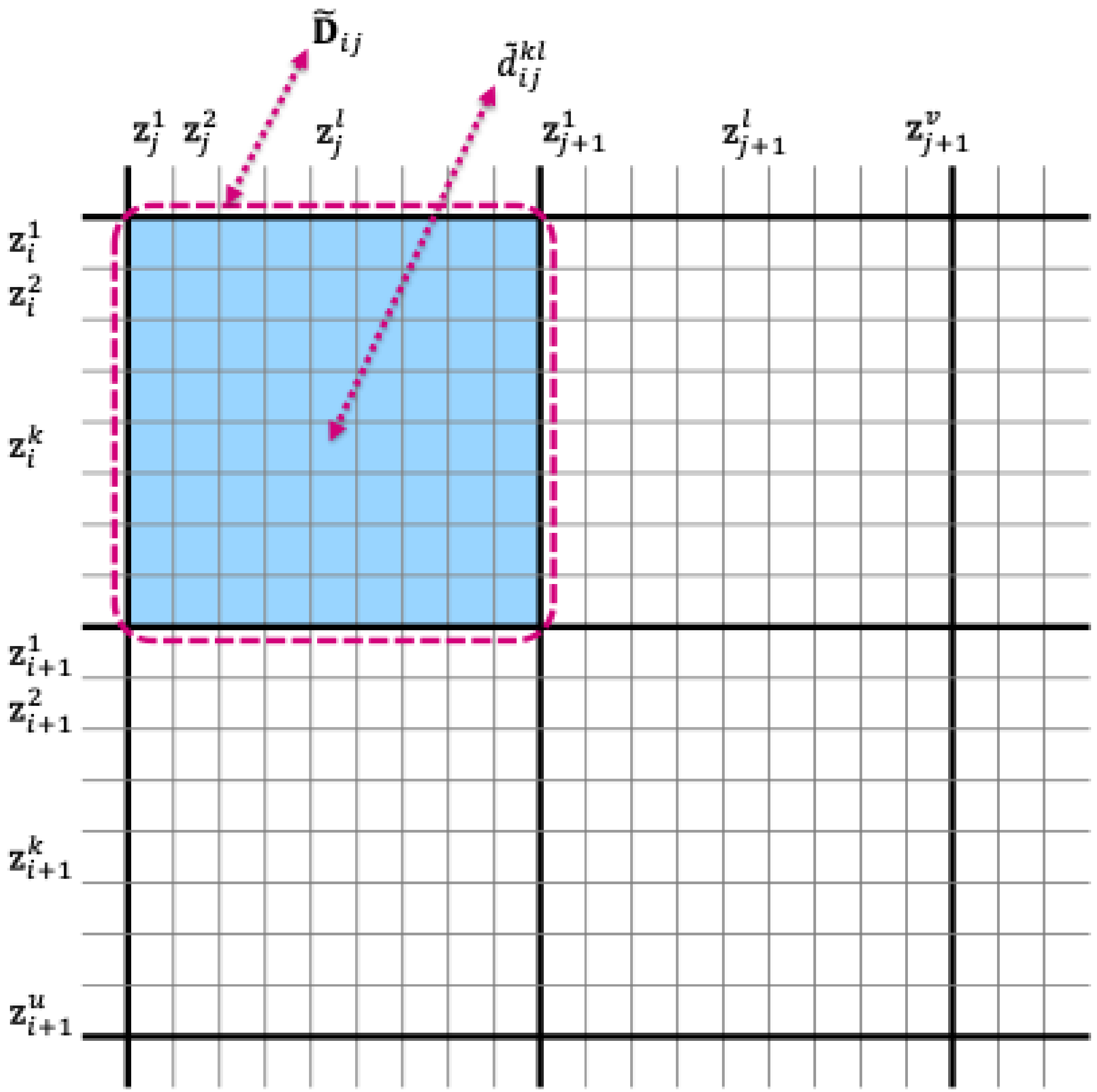


CLEWS: Reductions

Segment-based learning and matching:

- Reduction types: R_{mean}, $R_{\text{top-k}}$, R_{meanmin}, R_{min}

$$d_{ij} = \mathcal{R}_{\text{mean}}\left(\tilde{\mathbf{D}}_{ij}\right) = \frac{1}{uv} \sum_{\substack{1 \leq k \leq u \\ 1 \leq l \leq v}} \tilde{d}_{ij}^{kl}.$$

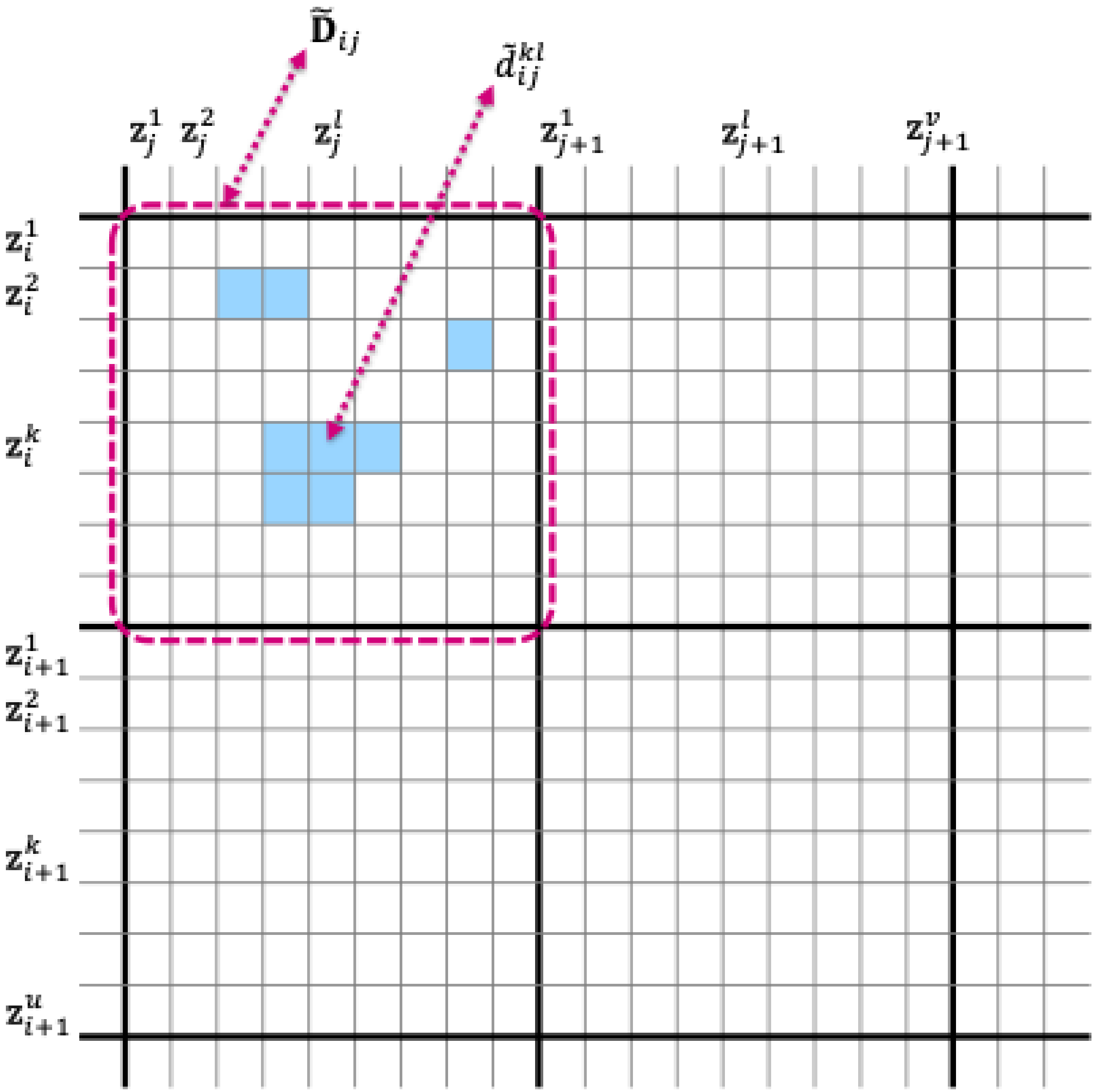


CLEWS: Reductions

Segment-based learning and matching:

- Reduction types: R_{mean}, $R_{\text{top-k}}$, R_{meanmin}, R_{min}

$$d_{ij} = \mathcal{R}_{\text{best-r}}\left(\tilde{\mathbf{D}}_{ij}\right) = \frac{1}{r} \sum_{1 \leq t \leq r} \operatorname{topr}\left(\tilde{\mathbf{D}}_{ij}\right)_t$$

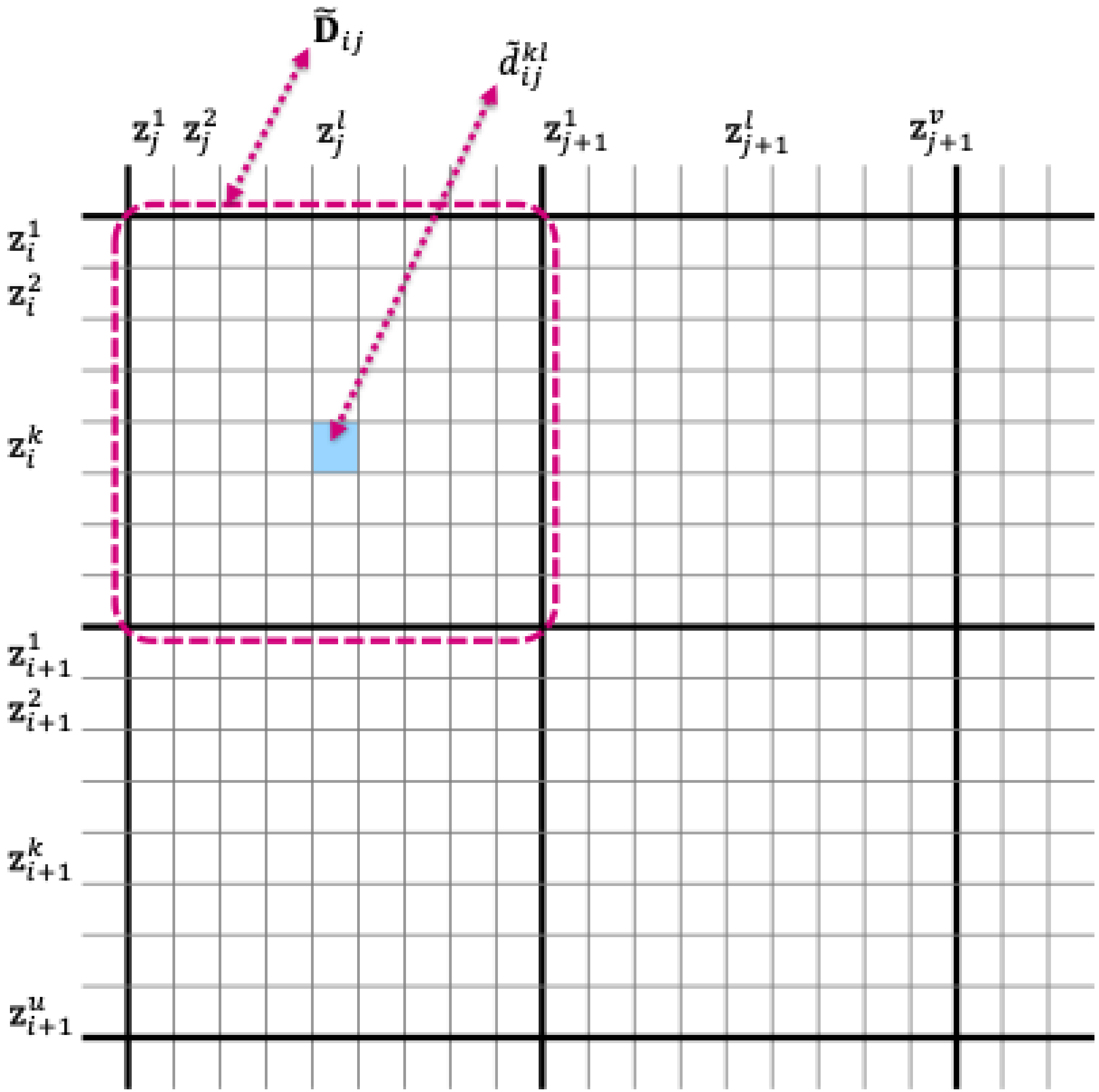


CLEWS: Reductions

Segment-based learning and matching:

- Reduction types: R_{mean}, $R_{\text{top-k}}$, R_{meanmin}, R_{min}

$$d_{ij} = \mathcal{R}_{\min}\left(\tilde{\mathbf{D}}_{ij}\right) = \min_{\substack{1\leq k\leq u \\ 1\leq l\leq v}} \tilde{d}_{ij}^{kl}.$$



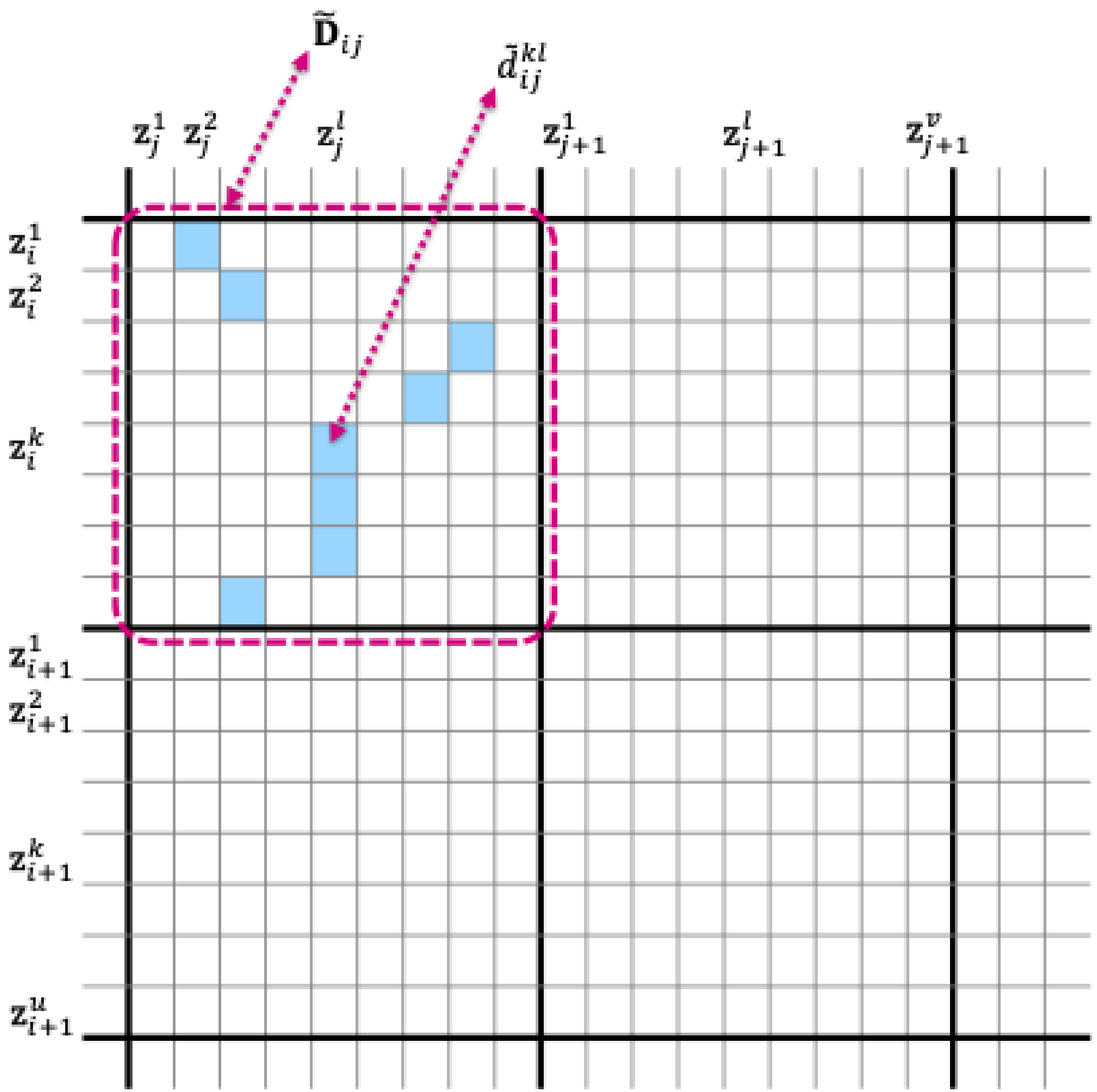
Sony AI

CLEWS: Reductions

Segment-based learning and matching:

- Reduction types: R_{mean}, $R_{\text{top-k}}$, R_{meanmin}, R_{min}

$$d_{ij} = \mathcal{R}_{\text{meanmin}}\left(\tilde{\mathbf{D}}_{ij}\right) = \frac{1}{u} \sum_{1 \leq k \leq u} \min_{1 \leq l \leq v} \tilde{d}_{ij}^{kl}.$$



Sony AI

CLEWS: Reductions

Segment-based learning and matching:

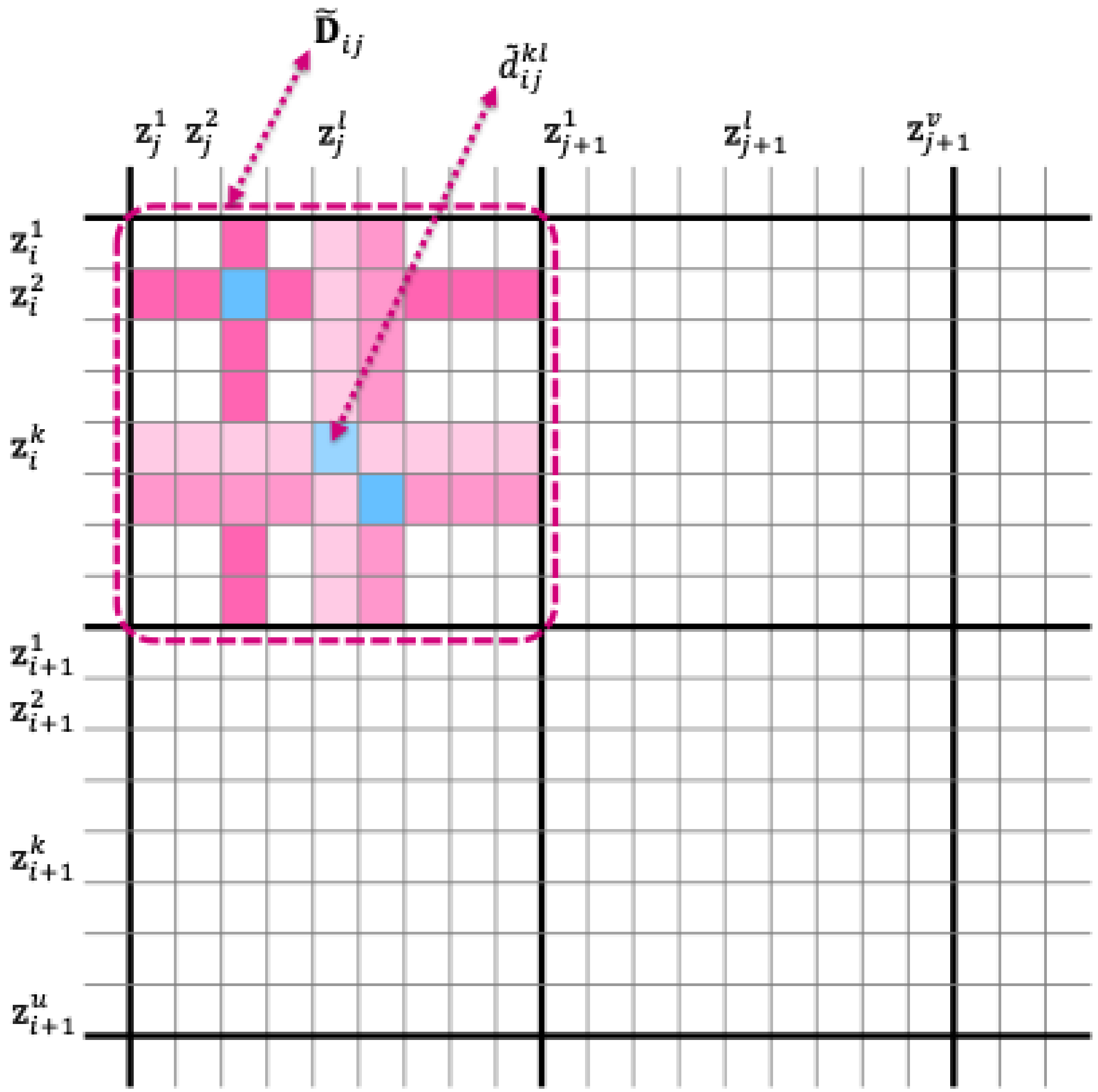
- Reduction types: R_{mean}, $R_{\text{top-k}}$, R_{meanmin}, R_{min}
- New reduction type: $R_{\text{bpwr-k}}$

$$d_{ij} = \mathcal{R}_{\text{bpwr-r}}\left(\tilde{\mathbf{D}}_{ij}\right) = \frac{1}{r}\sum_{1\leq q\leq r}\mathcal{R}_{\text{min}}\left(\tilde{\mathbf{D}}_{ij}^{(q)}\right) \qquad (1)$$

for $r \leq \min(u, v)$, with the recursion

$$\tilde{\mathbf{D}}_{ij}^{(q)} = \begin{cases} \tilde{\mathbf{D}}_{ij} & \text{for } q = 1, \\ \text{maskmin}\left(\tilde{\mathbf{D}}_{ij}^{(q-1)}\right) & \text{for } q > 1, \end{cases}$$

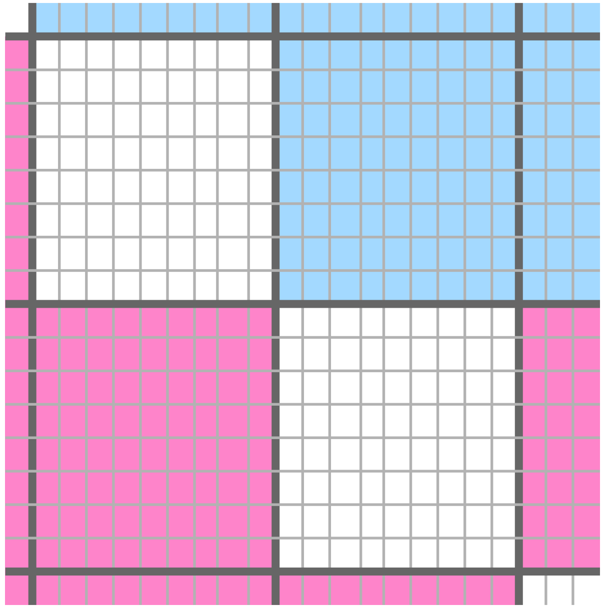
where $\text{maskmin}(\mathbf{D})$ is a function that masks the row and the column corresponding to the minimum element in $\mathbf{D}$, such



CLEWS: Reductions

Segment-based learning and matching:

- Reduction types: R_{mean}, $R_{\text{top-k}}$, R_{meanmin}, R_{min}
- New reduction type: $R_{\text{bpwr-k}}$

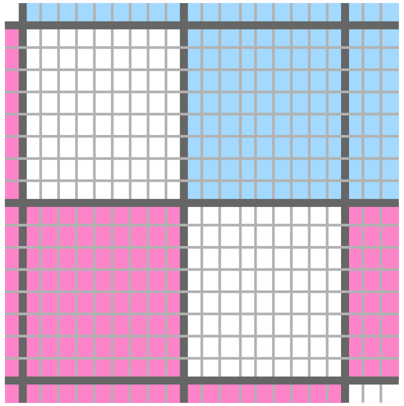


- Different reductions for positive and negative pairs:

$$\mathbf{D} = \mathbf{A} \odot \mathcal{R}^{+}(\tilde{\mathbf{D}}) + (\mathbf{1} - \mathbf{A}) \odot \mathcal{R}^{-}(\tilde{\mathbf{D}})$$

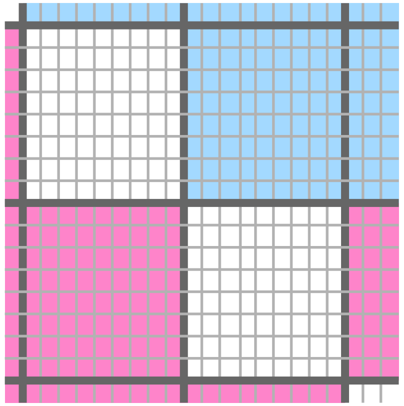
Sony AI

CLEWS: Loss



Contrastive loss. Starting from A&U.

CLEWS: Loss

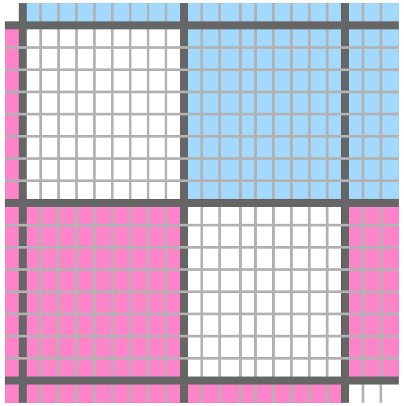


Contrastive loss. Starting from A&U.

- Decoupled: No overlap between positive and negative pairs.
- Change hyper-parameters: Fix/remove/add.
- “Comparable” gradients for positive & negative pairs.

$$\tilde{\mathcal{L}} = \frac{1}{|A^+|} \sum_{(i,j)\in A^+} d_{ij}^{2} + \cancel{\lambda} \log \left(\frac{1}{|A^-|} \sum_{(i,j)\in A^-} e^{-\gamma d_{ij}^2} \right), \quad (3)$$

CLEWS: Loss



Contrastive loss. Starting from A&U.

- Decoupled: No overlap between positive and negative pairs.
- Change hyper-parameters: Fix/remove/add.
- “Comparable” gradients for positive & negative pairs.
- Euclidean geometry *and* distance. Space geometry and geodesic distance should match.

$$\mathcal{L} = \frac{1}{|A^+|} \sum_{(i,j)\in A^+} d_{ij}^2 + \log\left(\varepsilon + \frac{1}{|A^-|} \sum_{(i,j)\in A^-} e^{-\gamma d_{ij}^2} \right)$$

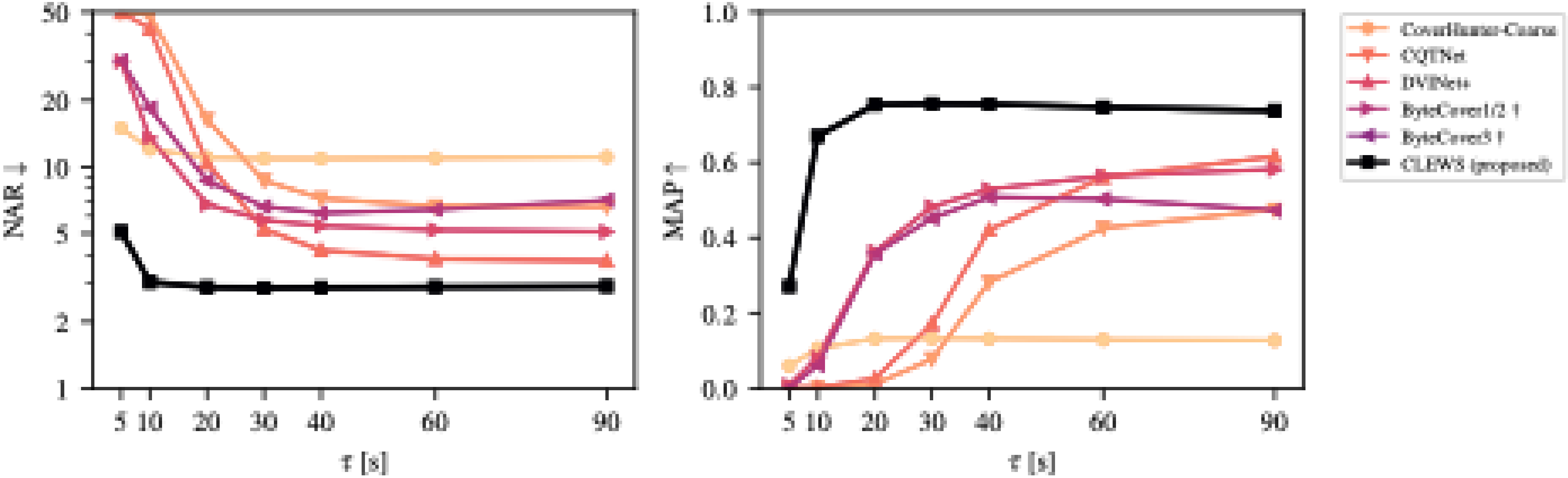
Sony AI

Results: Track-level

Table 2. Track-level evaluation and comparison with the state of the art. The symbol † denotes that it is our implementation.

APPROACH	DVI-TEST		SHS-TEST	
	NAR ↓	MAP ↑	NAR ↓	MAP ↑
COVERHUNTER-COARSE (LIU ET AL., 2023)	10.36 ± 0.07	0.157 ± 0.001	4.09 ± 0.17	0.491 ± 0.007
MOVE (YESILER ET AL., 2020A)	N/A	N/A	N/A	0.519
CQTNET (YU ET AL., 2020)	6.68 ± 0.07	0.493 ± 0.002	2.67 ± 0.16	0.677 ± 0.007
DVINET+ (ARAZ ET AL., 2024B)	3.69 ± 0.06	0.643 ± 0.002	2.39 ± 0.16	0.720 ± 0.007
LYRAC-NET (HU ET AL., 2022)	N/A	N/A	N/A	0.765
BYTECOVER3† (BASED ON DU ET AL., 2023)	5.64 ± 0.05	0.513 ± 0.002	1.91 ± 0.14	0.783 ± 0.006
BYTECOVER1/2† (BASED ON DU ET AL., 2022)	4.98 ± 0.06	0.595 ± 0.002	1.95 ± 0.14	0.813 ± 0.006
BYTECOVER3 (DU ET AL., 2023)	N/A	N/A	N/A	0.824
BYTECOVER3.5 (DU ET AL., 2024)	N/A	N/A	N/A	0.857
BYTECOVER2 (DU ET AL., 2022)	N/A	N/A	N/A	0.863
CLEWS (PROPOSED)	$\mathbf{2.70 \pm 0.05}$	$\mathbf{0.774 \pm 0.002}$	$\mathbf{1.27 \pm 0.12}$	$\mathbf{0.876 \pm 0.005}$

Results: Segment-level



(τ = Segment length)

Results: Reduction and Loss Ablations

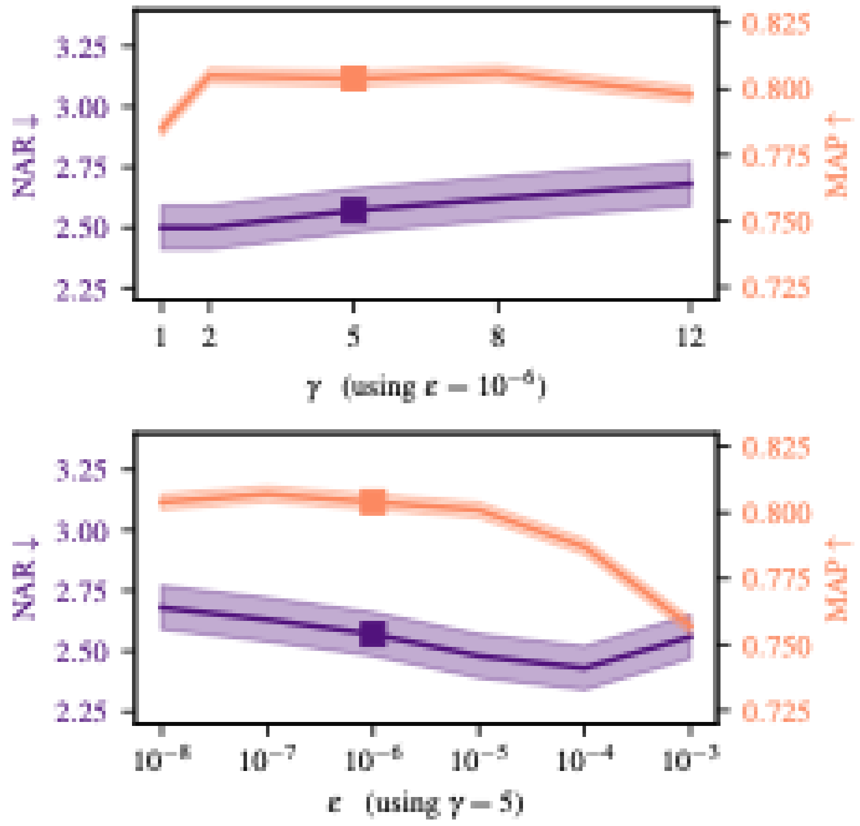
Table 3. Results on DVI-Valid for different positive $\mathcal{R}^+$ and negative $\mathcal{R}^-$ distance reductions. The default CLEWS reductions are $\mathcal{R}^+ = \mathcal{R}_{\text{bpwr-5}}$ and $\mathcal{R}^- = \mathcal{R}_{\text{min}}$.

$\mathcal{R}^+$	$\mathcal{R}^-$	NAR ↓	MAP ↑
CLEWS (PROPOSED)		2.57 ± 0.09	0.804 ± 0.003
$\mathcal{R}_{\text{bpwr-3}}$	$\mathcal{R}_{\text{min}}$	2.60 ± 0.09	**0.809 ± 0.003**
$\mathcal{R}_{\text{bpwr-8}}$	$\mathcal{R}_{\text{min}}$	**2.51 ± 0.09**	0.789 ± 0.003
$\mathcal{R}_{\text{meanmin}}$	$\mathcal{R}_{\text{min}}$	2.58 ± 0.09	0.798 ± 0.003
$\mathcal{R}_{\text{best-10}}$	$\mathcal{R}_{\text{min}}$	2.63 ± 0.09	0.788 ± 0.003
$\mathcal{R}_{\text{min}}$	$\mathcal{R}_{\text{min}}$	2.79 ± 0.09	0.799 ± 0.003
$\mathcal{R}_{\text{bpwr-5}}$	$\mathcal{R}_{\text{best-10}}$	2.82 ± 0.10	0.779 ± 0.003
$\mathcal{R}_{\text{bpwr-5}}$	$\mathcal{R}_{\text{bpwr-5}}$	2.88 ± 0.10	0.778 ± 0.003
$\mathcal{R}_{\text{bpwr-5}}$	$\mathcal{R}_{\text{meanmin}}$	4.95 ± 0.12	0.488 ± 0.004

Table 4. Results on DVI-Valid for different loss functions using the default CLEWS reductions of $\mathcal{R}^+ = \mathcal{R}_{\text{bpwr-5}}$ and $\mathcal{R}^- = \mathcal{R}_{\text{min}}$.

LOSS FUNCTION	NAR ↓	MAP ↑
CLEWS (PROPOSED)	**2.57 ± 0.09**	**0.804 ± 0.003**
NT-XENT	2.61 ± 0.09	0.732 ± 0.004
SUPCON	2.69 ± 0.09	0.676 ± 0.004
SIGLIP	2.79 ± 0.09	0.684 ± 0.004
TRIPLET	3.08 ± 0.11	0.717 ± 0.004
SUPCON-DECOUPLED	3.14 ± 0.11	0.739 ± 0.004
A&U-DECOUPLED	3.25 ± 0.11	0.620 ± 0.004
CLASSIFICATION XENT	8.91 ± 0.14	0.205 ± 0.003

Results: Hyper-parameters



Sony AI

Conclusion

- A state-of-the-art approach for musical version matching at the track level.
- Also breakthrough performance on musical version matching at the segment level.
- Based on two novel contributions:
  - Weak labeling → Segment reductions.
  - A&U loss → CLEWS loss (decoupling, hyperparameters, geometric considerations)
- Generality of the proposed concepts may make CLEWS applicable to further problems beyond music matching.