Generalizations of the Normalized Radon Cumulative Distribution Transform for Limited Data Recognition

Matthias Beckmann[1*], Robert Beinert[2*] and Jonas Bresch[2*]

[1]Fachbereich Mathematik, Universität Hamburg, Bundesstraße 55, Hamburg, 20146, Hamburg, Germany.

[2]Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, Berlin, 10623, Berlin, Germany.

*Corresponding author(s). E-mail(s): research@mbeckmann.de; beinert@math.tu-berlin.de; bresch@math.tu-berlin.de;

Abstract

The Radon cumulative distribution transform (R-CDT) exploits one-dimensional Wasserstein transport and the Radon transform to represent prominent features in images. It is closely related to the sliced Wasserstein distance and facilitates classification tasks, especially in the small data regime, like the recognition of watermarks in filigranology. Here, a typical issue is that the given data may be subject to affine transformations caused by the measuring process. To make the R-CDT invariant under arbitrary affine transformations, a two-step normalization of the R-CDT has been proposed in our earlier works. The aim of this paper is twofold. First, we propose a family of generalized normalizations to enhance flexibility for applications. Second, we study multi-dimensional and non-Euclidean settings by making use of generalized Radon transforms. We prove that our novel feature representations are invariant under certain transformations and allow for linear separation in feature space. Our theoretical results are supported by numerical experiments based on 2d images, 3d shapes and 3d rotation matrices, showing near perfect classification accuracies and clustering results.

Keywords: Radon-CDT, feature representation, image and shape classification, pattern recognition, small data regime

1 Introduction

Automated pattern recognition and classification play a central role in numerous applications and disciplines, be it in medical imaging, biometrics, or document analysis. Nowadays, in the big data regime, end-to-end deep neural networks provide the latest state of the art. In the small data regime, however, hand-crafted feature extractors and similarity measures still stand their ground.

In recent years, optimal transport-based techniques like the Wasserstein distance and variations of this have become popular tools for image comparison. At their core, the Wasserstein distances define metrics between probability measures on a common Polish space and find an optimal coupling [1–4]. Although being based on a linear program, the numerical calculation is challenging, initiating the development of entropic optimal transport [5], linear optimal transport [6–8], and the so-called sliced Wasserstein distance [9–12] to lower the computational burden. The definition of the Wasserstein distance can be extended

to ensure invariance under orthogonal transformations of the input measures, leading to the so-called Procrustes–Wasserstein distance [13, 14].

The optimal transport technique behind the Wasserstein distance can be further generalized to the so-called Gromov–Wasserstein [15] and embedded Wasserstein distance [16, 17], which both allow for the comparison of measures on different Polish spaces. These distances play a major role in shape and graph analysis since they are invariant under isometric transformations. In practice, they are calculated by solving costly quadratic programmes, limiting their applicability. As remedy, regularized [18], linear [19], and sliced [20–22] versions have been proposed.

Beyond similarity measures between images and shapes, optimal transport-based methods may be used to design feature extractors, being the focus of this paper. Ideally, a feature extractor transforms different classes to linearly separable subsets. This may, for instance, be achieved by the so-called Radon cumulative distribution transform (R-CDT) introduced in [23]. The R-CDT is based on one-dimensional optimal transport, which is generalized to two-dimensional data by applying the Radon transform, known from tomography [24, 25]. This approach shows great potential in many applications [26–28] and is closely related to the sliced Wasserstein distance. A similar approach for data on the sphere is studied in [29, 30], for multi-dimensional optimal transport maps in [8], and for optimal Gromov–Wasserstein transport maps in [19].

A central inspiration for this paper is the application of pattern recognition techniques in filigranology—the study of analogue watermarks. These play a central role in dating historical manuscripts as well as identifying scribes and papermills. For automatic classification, the main issue is the enormous number of classes with only few members per class. An end-to-end processing pipeline for thermograms of watermarks including an R-CDT-based classification is proposed in [31], where the authors report classification invariance with respect to translation and dilation of the watermark. In order to include other affine transformations caused, e.g., by unstandardized recording methods, in our previous works [32, 33], normalized R-CDT versions are proposed based on a two-step normalization scheme.

Contribution

This paper extends the max-normalized R-CDT ($_{\mathrm{m}}$NR-CDT) introduced at the SSVM'25 conference [32] by employing a more flexible normalization scheme and by generalizing the underlying Radon transform. Similar to [32], the unaltered goal is to design easy-to-implement feature extractors that are invariant under common transformations in the specific field of application. For instance, the watermark recognition task in filigranology requires feature extractors that are invariant under affine image transformations like rotation, shifting, shearing, scaling. Since this paper is an extension of [32] and thus substantially based on this, the first two contributions remain:

- the proposal of a new two-step normalization procedure for the R-CDT yielding the $_{\mathrm{m}}$NR-CDT feature extractor, which is invariant under arbitrary affine transformations,
- a rigorous study of the separability properties of the $_{\mathrm{m}}$NR-CDT, especially, affine classes become linear separable in $_{\mathrm{m}}$NR-CDT space,

The robustness of the original $_{\mathrm{m}}$NR-CDT under non-affine deformations is studied in [33], which is not the focus of this paper. In difference to [32, 33], this extended version contains the following additional contributions:

- a flexible final normalization for the NR-CDT yielding the new $_{h}$NR-CDT and enabling the usage of further angular informations for instance based on the total variation,
- the replacement of the underlying 2d Radon transform by the generalized Radon transform extending the $_{h}$NR-CDT to the multidimensional setting, the circular Radon transform or the Radon transform on the 3d rotation group SO(3),
- a rigours discussion about the linear separability properties, which provably carry over to all new extended NR-CDT variants,
- an extension of the image classification experiments in [32], by considering image clustering, 3d shape recognition, and classification of SO(3) point clouds.

Outline

The first part of this paper revisit the SSVM proceeding [32] where the $_{\mathrm{m}}$NR-CDT is originally proposed. To this end, we first generalize the classical Radon transform to measures in § 2 and, thereon, restate the definition of the $_{\mathrm{m}}$NR-CDT in § 3. The extension of the final normalization step is, especially, presented in § 3.4 yielding the new $_h$NR-CDT. The linear separability of affinely transformed measure classes in $_{\mathrm{m}}$NR-CDT and $_h$NR-CDT space is proven in Theorem 1 and 2. The novel generalization of the $_h$NR-CDT to arbitrary domains beyond the two-dimensional setting is introduced in § 4. The focus here lies on the extension to $\mathbb{R}^d$ related to the multidimensional and circular Radon transform as well as to the 3d rotation group. The separability properties for these cases are reported in Theorem 3–5. Our theoretical findings are supported by proof-of-concept experiments in § 5 showing significant improvements of the classification accuracy by the proposed normalizations, especially, in the small data regime.

2 Radon Transform

The main idea behind the classical Radon transform [25] is to integrate a given bivariate function along all parallel lines pointing in a certain direction. This integral transform can also be interpreted as projection of the given function onto the line with orthogonal orientation. In the following, we briefly review the classical Radon transform for functions and generalize the concept to measures. Finally, we study the effect of affine transformations on the Radon transform, which is crucial to solve the classification task at hand.

2.1 Radon Transform of Functions

Depending on $\boldsymbol{\theta} \in \mathbb{S}_1 \coloneqq \{\mathbf{x} \in \mathbb{R}^2 \mid \|\mathbf{x}\| = 1\}$, we introduce the *slicing operator* $S_{\boldsymbol{\theta}} \colon \mathbb{R}^2 \to \mathbb{R}$ by

$$S_{\boldsymbol{\theta}}(\mathbf{x}) \coloneqq \langle \mathbf{x}, \boldsymbol{\theta} \rangle, \quad \mathbf{x} \in \mathbb{R}^2.$$

Its preimages $S_{\boldsymbol{\theta}}^{-1}(t)$, $t \in \mathbb{R}$, are the lines $\ell_{t,\boldsymbol{\theta}}$ in direction $\boldsymbol{\theta}^\perp \coloneqq (\theta_2, -\theta_1)^\top \in \mathbb{S}_1$ with distance t to the origin. More precisely, we have

$$\ell_{t,\boldsymbol{\theta}} \coloneqq S_{\boldsymbol{\theta}}^{-1}(t) = \{t\boldsymbol{\theta} + \tau\boldsymbol{\theta}^\perp \mid \tau \in \mathbb{R}\} \subset \mathbb{R}^2.$$

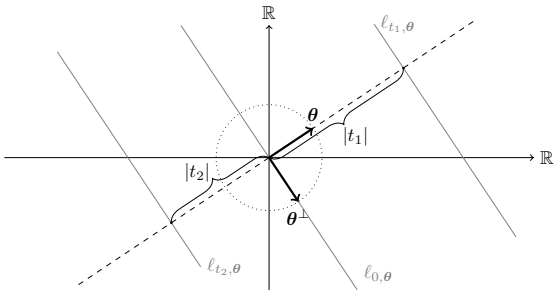

Fig. 1 Illustration of the bivariate Radon transform with distance $t \in \mathbb{R}$ and normal direction $\boldsymbol{\theta} \in \mathbb{S}_1$. The given function is integrated along the lines $\ell_{t,\boldsymbol{\theta}}$.

Using the bijection $\varphi_{\boldsymbol{\theta}} \colon \mathbb{R}^2 \to \mathbb{R}^2$ defined as $\varphi_{\boldsymbol{\theta}}(t,\tau) \coloneqq t\,\boldsymbol{\theta} + \tau\,\boldsymbol{\theta}^\perp$, whose inverse is given by $\varphi_{\boldsymbol{\theta}}(\mathbf{x})^{-1} = (\langle \mathbf{x}, \boldsymbol{\theta} \rangle, \langle \mathbf{x}, \boldsymbol{\theta}^\perp \rangle)$, we parameterize $\ell_{t,\boldsymbol{\theta}}$ via $\tau \mapsto \varphi_{\boldsymbol{\theta}}(t,\tau)$.

For $f \in L^1(\mathbb{R}^2)$, we define its *Radon transform* $\mathcal{R}[f] \colon \mathbb{R} \times \mathbb{S}_1 \to \mathbb{R}$ as the line integral

$$\mathcal{R}[f](t,\boldsymbol{\theta}) \coloneqq \int_{\ell_{t,\boldsymbol{\theta}}} f(s)\,\mathrm{d}s, \quad (t,\boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{S}_1,$$

where $\mathrm{d}s$ denotes the arc length element of $\ell_{t,\boldsymbol{\theta}}$. This defines the *Radon operator* $\mathcal{R} \colon L^1(\mathbb{R}^2) \to L^1(\mathbb{R} \times \mathbb{S}_1)$. For fixed $\boldsymbol{\theta} \in \mathbb{S}_1$, we set $\mathcal{R}_{\boldsymbol{\theta}} \coloneqq \mathcal{R}(\cdot, \boldsymbol{\theta})$, which is referred to as the *restricted Radon operator* $\mathcal{R}_{\boldsymbol{\theta}} \colon L^1(\mathbb{R}^2) \to L^1(\mathbb{R})$. The action of the Radon operator is illustrated in Figure 1.

The Radon transform is also well-defined for all $f \in L^p(\mathbb{R}^2)$ with $p \geq 1$ and $\operatorname{supp}(f) \subseteq \mathbb{B}_2 \coloneqq \{\mathbf{x} \in \mathbb{R}^2 \mid \|\mathbf{x}\| \leq 1\}$, in which case $\mathcal{R}[f] \in L^p(\mathbb{R} \times \mathbb{S}_1)$ with $\operatorname{supp}(\mathcal{R}[f]) \subseteq \mathbb{I} \times \mathbb{S}_1$, where $\mathbb{I} \coloneqq [-1,1]$. According to [25], the adjoint operator $\mathcal{R}^* \colon L^\infty(\mathbb{R} \times \mathbb{S}_1) \to L^\infty(\mathbb{R}^2)$ of the Radon transform $\mathcal{R} \colon L^1(\mathbb{R}^2) \to L^1(\mathbb{R} \times \mathbb{S}_1)$ is given by the *back projection*

$$\mathcal{R}^*[g](\mathbf{x}) \coloneqq \frac{1}{2\pi} \int_{\mathbb{S}_1} g(S_{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\theta})\,\mathrm{d}\sigma_{\mathbb{S}_1}(\boldsymbol{\theta}), \quad \mathbf{x} \in \mathbb{R}^2,$$

where $\sigma_{\mathbb{S}_1}$ denotes the surface measure on $\mathbb{S}_1$.

2.2 Radon Transform of Measures

The concept of the Radon transform is now translated to signed, regular, finite measures $\mu \in \mathcal{M}(\mathbb{R}^2)$. For a fixed direction $\boldsymbol{\theta} \in \mathbb{S}_1$, we generalize the *restricted Radon transform* $\mathcal{R}_{\boldsymbol{\theta}}$ to measures by

setting

$$\mathcal{R}_{\boldsymbol{\theta}} \colon \mathcal{M}(\mathbb{R}^2) \to \mathcal{M}(\mathbb{R}), \quad \mu \mapsto (S_{\boldsymbol{\theta}})_{\#}\mu = \mu \circ S_{\boldsymbol{\theta}}^{-1},$$

which corresponds to the integration along $\ell_{t,\boldsymbol{\theta}}$. Note that $\mathcal{R}_{\boldsymbol{\theta}}[\mu](\mathbb{R}) = \mu(\mathbb{R}^2)$ for all $\boldsymbol{\theta} \in \mathbb{S}_1$ and, thus, the mass of μ is preserved by $\mathcal{R}_{\boldsymbol{\theta}}$. In measure theory, $\mathcal{R}_{\boldsymbol{\theta}}$ can be considered as a disintegration family. Heuristically, we may generalize the Radon transform by integrating $\mathcal{R}_{\boldsymbol{\theta}}$ along $\boldsymbol{\theta} \in \mathbb{S}_1$. Therefore, we define the *Radon transform* $\mathcal{R} \colon \mathcal{M}(\mathbb{R}^2) \to \mathcal{M}(\mathbb{R} \times \mathbb{S}_1)$ via

$$\mathcal{R}[\mu] \coloneqq \mathcal{I}_{\#}[\mu \times u_{\mathbb{S}_1}] \tag{1}$$

with $\mathcal{I}(\mathbf{x}, \boldsymbol{\theta}) \coloneqq (S_{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\theta})$ for $(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^2 \times \mathbb{S}_1$. Here $\mu \times u_{\mathbb{S}_1}$ denotes the product measure between the given μ and the uniform measure $u_{\mathbb{S}_1} \coloneqq \sigma_{\mathbb{S}_1}/2\pi$ on $\mathbb{S}_1$.

Proposition 1 *Let $\mu \in \mathcal{M}(\mathbb{R}^2)$. Then, $\mathcal{R}[\mu]$ can be disintegrated into the family $\mathcal{R}_{\boldsymbol{\theta}}[\mu]$ with respect to $u_{\mathbb{S}_1}$, i.e., for all continuous $g \in C_0(\mathbb{R} \times \mathbb{S}_1)$ vanishing at infinity, we have*

$$\langle \mathcal{R}[\mu], g \rangle = \int_{\mathbb{S}_1} \langle \mathcal{R}_{\boldsymbol{\theta}}[\mu], g(\cdot, \boldsymbol{\theta}) \rangle \, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}).$$

Proof By definition in (1), we obtain

$$\begin{aligned}
\langle \mathcal{R}[\mu], g \rangle &= \int_{\mathbb{R} \times \mathbb{S}_1} g(t, \boldsymbol{\theta}) \, \mathrm{d}\mathcal{I}_{\#}[\mu \times u_{\mathbb{S}_1}](t, \boldsymbol{\theta}) \\
&= \int_{\mathbb{S}_1} \int_{\mathbb{R}^2} g(S_{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\theta}) \, \mathrm{d}\mu(\mathbf{x}) \, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) \\
&= \int_{\mathbb{S}_1} \int_{\mathbb{R}^2} g(t, \boldsymbol{\theta}) \, \mathrm{d}[(S_{\boldsymbol{\theta}})_{\#}\mu](t) \, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) \\
&= \int_{\mathbb{S}_1} \langle \mathcal{R}_{\boldsymbol{\theta}}[\mu], g(\cdot, \boldsymbol{\theta}) \rangle \, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta})
\end{aligned}$$

using Fubini's theorem. □

One can find the measure-valued Radon transform $\mathcal{R} \colon \mathcal{M}(\mathbb{R}^2) \to \mathcal{M}(\mathbb{R} \times \mathbb{S}_1)$ as the adjoint of the function-valued adjoint $\mathcal{R}^* \colon L^\infty(\mathbb{R} \times \mathbb{S}_1) \to L^\infty(\mathbb{R}^2)$, similar to the case of distributions with compact support, cf. [24].

Proposition 2 *The Radon transform of $\mu \in \mathcal{M}(\mathbb{R}^2)$ satisfies*

$$\langle \mathcal{R}[\mu], g \rangle = \langle \mu, \mathcal{R}^*[g] \rangle \quad \forall\, g \in L^\infty(\mathbb{R} \times \mathbb{S}_1).$$

Proof For all $\mu \in \mathcal{M}(\mathbb{R}^2)$ and $g \in L^\infty(\mathbb{R} \times \mathbb{S}_1)$, applying Fubini's theorem gives

$$\begin{aligned}
\langle \mathcal{R}[\mu], g \rangle &= \int_{\mathbb{R} \times \mathbb{S}_1} g(t, \boldsymbol{\theta}) \, \mathrm{d}\mathcal{I}_{\#}[\mu \times u_{\mathbb{S}_1}](t, \boldsymbol{\theta}) \\
&= \int_{\mathbb{R}^2} \int_{\mathbb{S}_1} g(S_{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\theta}) \, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) \, \mathrm{d}\mu(\mathbf{x}).
\end{aligned}$$

□

Note that, for $f \in L^1(\mathbb{R}^2)$ and the Lebesgue measure $\lambda_{\mathbb{R}^2}$ on $\mathbb{R}^2$, the Radon transform satisfies

$$\mathcal{R}[f \, \lambda_{\mathbb{R}^2}] = \mathcal{R}[f] \, \sigma_{\mathbb{R} \times \mathbb{S}_1},$$

where $\sigma_{\mathbb{R} \times \mathbb{S}_1}$ denotes the surface measure on $\mathbb{R} \times \mathbb{S}_1$. In particular, the Radon transform of an absolutely continuous measure is again absolutely continuous.

2.3 Radon Transform of Affine Transformations

We now consider the Radon transform of an affinely transformed finite measure $\mu \in \mathcal{M}(\mathbb{R}^2)$. To this end, let $\mathbf{A} \in \mathrm{GL}(2)$ and $\mathbf{y} \in \mathbb{R}^2$, i.e., $\mathbf{A}$ is contained in the general linear group GL of regular matrices. We define $\mu_{\mathbf{A},\mathbf{y}} \in \mathcal{M}(\mathbb{R}^2)$ via

$$\mu_{\mathbf{A},\mathbf{y}} \coloneqq (\mathbf{A} \cdot + \mathbf{y})_{\#}\mu = \mu \circ (\mathbf{A}^{-1}(\cdot - \mathbf{y})). \tag{2}$$

Proposition 3 *For any $\boldsymbol{\theta} \in \mathbb{S}_1$, the restricted Radon transform satisfies*

$$\begin{aligned}
\mathcal{R}_{\boldsymbol{\theta}}[\mu_{\mathbf{A},\mathbf{y}}] &= (\|\mathbf{A}^\top \boldsymbol{\theta}\| \cdot + \langle \mathbf{y}, \boldsymbol{\theta} \rangle)_{\#} \mathcal{R}_{\frac{\mathbf{A}^\top \boldsymbol{\theta}}{\|\mathbf{A}^\top \boldsymbol{\theta}\|}}[\mu] \\
&= \mathcal{R}_{\frac{\mathbf{A}^\top \boldsymbol{\theta}}{\|\mathbf{A}^\top \boldsymbol{\theta}\|}}[\mu] \circ \left(\tfrac{\cdot - \langle \mathbf{y}, \boldsymbol{\theta} \rangle}{\|\mathbf{A}^\top \boldsymbol{\theta}\|} \right).
\end{aligned}$$

Proof Direct calculations yield

$$\begin{aligned}
\mathcal{R}_{\boldsymbol{\theta}}[\mu_{\mathbf{A},\mathbf{y}}] &= (S_{\boldsymbol{\theta}})_{\#}[(\mathbf{A} \cdot + \mathbf{y})_{\#}\mu] \\
&= (\langle \mathbf{A} \cdot + \mathbf{y}, \boldsymbol{\theta} \rangle)_{\#}\mu \\
&= (\langle \cdot, \mathbf{A}^\top \boldsymbol{\theta} \rangle + \langle \mathbf{y}, \boldsymbol{\theta} \rangle)_{\#}\mu \\
&= (\|\mathbf{A}^\top \boldsymbol{\theta}\| \langle \cdot, \tfrac{\mathbf{A}^\top \boldsymbol{\theta}}{\|\mathbf{A}^\top \boldsymbol{\theta}\|} \rangle + \langle \mathbf{y}, \boldsymbol{\theta} \rangle)_{\#}\mu \\
&= (\|\mathbf{A}^\top \boldsymbol{\theta}\| \cdot + \langle \mathbf{y}, \boldsymbol{\theta} \rangle)_{\#} \mathcal{R}_{\frac{\mathbf{A}^\top \boldsymbol{\theta}}{\|\mathbf{A}^\top \boldsymbol{\theta}\|}}[\mu],
\end{aligned}$$

and the proof is complete. □

The effect of common affine transformations on the Radon transform is given in Table 1. In order to describe the deformation with respect to $\boldsymbol{\theta}$, we over-parameterize the unit circle $\mathbb{S}_1$ via $\boldsymbol{\theta}(\vartheta) \coloneqq$

transformation	$\mathbf{A}$	$\mathbf{y}$	$\mathcal{R}_{\boldsymbol{\theta}(\vartheta)}[\mu_{\mathbf{A},\mathbf{y}}]$, $\vartheta \in (-\frac{\pi}{2}, \frac{\pi}{2})$
translation	$\mathbf{I}$	$\mathbb{R}^2$	$\mathcal{R}_{\boldsymbol{\theta}(\vartheta)}[\mu] \circ (\cdot - \langle \mathbf{y}, \boldsymbol{\theta}(\vartheta) \rangle)$
rotation	$\left(\begin{smallmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{smallmatrix}\right)$	$\mathbf{0}$	$\mathcal{R}_{\boldsymbol{\theta}(\vartheta-\varphi)}[\mu]$
reflection	$\left(\begin{smallmatrix} \cos(\varphi) & \sin(\varphi) \\ \sin(\varphi) & -\cos(\varphi) \end{smallmatrix}\right)$	$\mathbf{0}$	$\mathcal{R}_{\boldsymbol{\theta}(\varphi-\vartheta)}[\mu]$
anisotropic scaling	$\left(\begin{smallmatrix} a & 0 \\ 0 & b \end{smallmatrix}\right)$	$\mathbf{0}$	$\mathcal{R}_{\boldsymbol{\theta}(\arctan(\frac{b}{a}\tan(\vartheta)))}[\mu] \circ ([a^2\cos^2(\vartheta) + b^2\sin^2(\vartheta)]^{-1/2}\cdot)$
vertical shear	$\left(\begin{smallmatrix} 1 & 0 \\ c & 1 \end{smallmatrix}\right)$	$\mathbf{0}$	$\mathcal{R}_{\boldsymbol{\theta}(\arctan(c+\tan(\vartheta)))}[\mu] \circ ([1 + c^2\cos^2(\vartheta) + c\sin(2\vartheta)]^{-1/2}\cdot)$

Table 1 Summary of common transformations for $\mu \in \mathcal{M}(\mathbb{R}^2)$ with $a, b > 0$ and $c, \varphi \in \mathbb{R}$. The unit circle is parameterized by $\boldsymbol{\theta}(\vartheta) \coloneqq (\cos(\vartheta), \sin(\vartheta))^\top$. The Radon transform for the left half of $\mathbb{S}_1$ follows by symmetry.

$(\cos(\vartheta), \sin(\vartheta))^\top$, $\vartheta \in \mathbb{R}$. As by Proposition 3, an affine transformation essentially causes a translation and dilation of the transformed measure together with a non-affine remapping in $\boldsymbol{\theta}$.

3 Optimal Transport-Based Transforms

The aim of the following is to introduce an image distance that is unaware of affine transformations. Methodologically, we rely on the *Radon cumulative distribution transform* (R-CDT) introduced in [23], which allows to utilize the fast-to-compute, one-dimensional Wasserstein distance in the context of image processing due to a Radon-based slicing technique. As the R-CDT is not invariant under affine transformation by itself, we propose a two-step normalization scheme, which is essentially grounded on our observations regarding the Radon transform under affine transformations in § 2.3. Finally, we study the linear separability of affinely transformed image classes by our novel normalized R-CDT.

3.1 R-CDT for Measures

The R-CDT traces back to Kolouri et al. [23] and transforms smooth, bivariate density functions. In contrast to [23], we introduce the concept for arbitrary probability measures, similar to [27]. In a first step, we consider probability measures $\mathcal{P}(\mathbb{R}) \subset \mathcal{M}(\mathbb{R})$ defined on the real line. For $\mu \in \mathcal{P}(\mathbb{R})$, the *cumulative distribution function* $F_\mu \colon \mathbb{R} \to [0,1]$ is given by $F_\mu(t) \coloneqq \mu((-\infty, t])$, $t \in \mathbb{R}$. Its generalized inverse, known as *quantile function*, reads as

$$F_\mu^{[-1]}(t) \coloneqq \inf\{s \in \mathbb{R} \mid F_\mu(s) > t\}, \quad t \in \mathbb{R}.$$

Based on a reference measure $\rho \in \mathcal{P}(\mathbb{R})$ that does not give mass to atoms, e.g., the uniform distribution $u_{[0,1]}$ on $[0,1]$, we define the *cumulative distribution transform* $\widehat{\mu} \colon \mathbb{R} \to \mathbb{R}$, in short CDT, via

$$\widehat{\mu} \coloneqq F_\mu^{[-1]} \circ F_\rho.$$

For any convex cost function $c \colon \mathbb{R} \to [0, \infty)$, the CDT (with respect to ρ) solves the Monge–Kantorovich transportation problem [2], i.e.,

$$\widehat{\mu} = \operatorname*{arg\,min}_{T_\# \rho = \mu} \int_{\mathbb{R}} c(s - T(s)) \, \mathrm{d}\rho(s),$$

where the minimum is taken over all measurable functions $T \colon \mathbb{R} \to \mathbb{R}$. In other words, $\widehat{\mu} \colon \mathbb{R} \to \mathbb{R}$ is an optimal Monge map transporting ρ to μ while minimizing the cost. If $\mu \in \mathcal{P}_2(\mathbb{R})$, i.e., μ has finite 2nd moment, then $\widehat{\mu}$ is square integrable with respect to ρ, i.e., $\widehat{\mu} \in L^2_\rho(\mathbb{R})$. Moreover, for $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$, the norm distance

$$\|\widehat{\mu} - \widehat{\nu}\|_\rho \coloneqq \Big(\int_{\mathbb{R}} |\widehat{\mu}(t) - \widehat{\nu}(t)|^2 \, \mathrm{d}\rho(t) \Big)^{\frac{1}{2}}$$

equals the well-established Wasserstein-2 distance [2].

To deal with a probability measure $\mu \in \mathcal{P}(\mathbb{R}^2)$ defined on the plane, we first determine the Radon transform $\mathcal{R}[\mu] \in \mathcal{M}(\mathbb{R} \times \mathbb{S}_1)$ with its disintegration family $\{\mathcal{R}_{\boldsymbol{\theta}}[\mu] \in \mathcal{P}(\mathbb{R}) \mid \boldsymbol{\theta} \in \mathbb{S}_1\}$. Then, for each fixed $\boldsymbol{\theta} \in \mathbb{S}_1$, we consider the CDT $\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]$ (with respect to the same reference measure $\rho \in \mathcal{P}(\mathbb{R})$ for all $\boldsymbol{\theta} \in \mathbb{S}_1$) of the Radon projection $\mathcal{R}_{\boldsymbol{\theta}}[\mu]$, yielding the *R-CDT* $\widehat{\mathcal{R}}[\mu] \colon \mathbb{R} \times \mathbb{S}_1 \to \mathbb{R}$ of μ via

$$\widehat{\mathcal{R}}[\mu](t, \boldsymbol{\theta}) \coloneqq \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t), \quad (t, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{S}_1.$$

If $\mu \in \mathcal{P}_2(\mathbb{R}^2)$, then the Radon projection $\mathcal{R}_{\boldsymbol{\theta}}[\mu] \in \mathcal{P}_2(\mathbb{R})$ has finite 2nd moment as well. Consequently, $\widehat{\mathcal{R}}[\mu] \in L^2_{\rho \times u_{\mathbb{S}_1}}(\mathbb{R} \times \mathbb{S}_1)$. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^2)$, the norm distance

$$\|\widehat{\mathcal{R}}[\mu] - \widehat{\mathcal{R}}[\nu]\|_{\rho \times u_{\mathbb{S}_1}} \coloneqq \Big(\int_{\mathbb{S}_1} \int_{\mathbb{R}} |\widehat{\mathcal{R}}[\mu](t, \boldsymbol{\theta}) - \widehat{\mathcal{R}}[\nu](t, \boldsymbol{\theta})|^2 \, \mathrm{d}\rho(t) \, \mathrm{d}u_{\mathbb{S}_1}(\boldsymbol{\theta}) \Big)^{\frac{1}{2}}$$

is also called sliced Wasserstein-2 distance in the literature [9].

3.2 Normalized R-CDT

The R-CDT is by itself not invariant under affine transformations, which emerge in various applications. More precisely, the R-CDT inherits the behavior of the Radon transform observed in § 2.3. Notice that the translation and dilation of $\mathcal{R}_{\boldsymbol{\theta}}[\mu]$ causes a horizontal shift (addition of a constant) and a scaling (multiplication with a constant) of $\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]$, respectively. In the first normalization step, we revert these effects by ensuring zero mean and unit standard deviation of the R-CDT projection. More precisely, we define the *normalized R-CDT* (NR-CDT) $\mathcal{N}[\mu] \colon \mathbb{R} \times \mathbb{S}_1 \to \mathbb{R}$ of $\mu \in \mathcal{P}_2(\mathbb{R}^2)$ via

$$\mathcal{N}[\mu](t, \boldsymbol{\theta}) \coloneqq \mathcal{N}_{\boldsymbol{\theta}}[\mu](t), \quad (t, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{S}_1,$$

with

$$\mathcal{N}_{\boldsymbol{\theta}}[\mu](t) \coloneqq \frac{\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t) - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])}{\operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])},$$

where, for $g \in \mathrm{L}^2_\rho(\mathbb{R})$,

$$\operatorname{mean}(g) \coloneqq \int_{\mathbb{R}} g(s) \, \mathrm{d}\rho(s)$$

and

$$\operatorname{std}(g) \coloneqq \Big(\int_{\mathbb{R}} |g(s) - \operatorname{mean}(g)|^2 \, \mathrm{d}\rho(s) \Big)^{\frac{1}{2}}.$$

To ensure that the NR-CDT is well defined, we have to guarantee that the standard deviation of the R-CDT projection does not vanish. For this, we restrict ourselves to measures whose supports are not contained in a straight line. More precisely, we consider the class

$$\mathcal{P}^*_{\mathrm{c}}(\mathbb{R}^2) \coloneqq \{\mu \in \mathcal{P}(\mathbb{R}^2) \,|\, \operatorname{supp}(\mu) \subset\subset \mathbb{R}^2 \wedge \dim(\operatorname{supp}(\mu)) > 1\}$$

satisfying

$$\mathcal{P}^*_{\mathrm{c}}(\mathbb{R}^2) \subset \mathcal{P}_2(\mathbb{R}^2).$$

Here, $\subset\subset$ denotes a compact subset, and dim the dimension of the affine hull. For these, the standard deviation of the restricted Radon transform is bounded away from zero and cannot vanish.

Proposition 4 *Let $\mu \in \mathcal{P}^*_{\mathrm{c}}(\mathbb{R}^2)$. Then, there exists a constant $c > 0$ such that*

$$\operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) \geq c \quad \forall \boldsymbol{\theta} \in \mathbb{S}_1.$$

For the proof, we first show the following continuity.

Lemma 1 *For fixed $\mu \in \mathcal{P}^*_{\mathrm{c}}(\mathbb{R}^2)$, the functions $\boldsymbol{\theta} \in \mathbb{S}_1 \mapsto \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) \in \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{S}_1 \mapsto \operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) \in \mathbb{R}_{\geq 0}$ are continuous.*

Proof We rewrite the mean as

$$\begin{aligned} \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) &= \int_{\mathbb{R}} \widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t) \, \mathrm{d}\rho(t) \\ &= \int_{\mathbb{R}} t \, \mathrm{d}\mathcal{R}_{\boldsymbol{\theta}}[\mu](t) \\ &= \int_{\mathbb{R}^2} \langle \mathbf{x}, \boldsymbol{\theta} \rangle \, \mathrm{d}\mu(\mathbf{x}). \end{aligned}$$

Since the integrand is continuous in $\boldsymbol{\theta}$ and uniformly bounded by $|\langle \cdot, \boldsymbol{\theta} \rangle| \leq \|\cdot\|$, the dominated convergence yields the assertion. Analogously, we have

$$\begin{aligned} \operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) &= \Big(\int_{\mathbb{R}} |\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t) - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])|^2 \, \mathrm{d}\rho(t) \Big)^{\frac{1}{2}} \\ &= \Big(\int_{\mathbb{R}^2} |\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])|^2 \, \mathrm{d}\mu(\mathbf{x}) \Big)^{\frac{1}{2}}. \end{aligned}$$

The integrand is again continuous in $\boldsymbol{\theta}$ and uniformly bounded by

$$|\langle \cdot, \boldsymbol{\theta} \rangle - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])|^2 \leq 2\|\cdot\|^2 + 2 \max_{\boldsymbol{\theta} \in \mathbb{S}_1} (\operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]))^2;$$

thus, the standard deviation is continuous by dominated convergence. □

Proof of Proposition 4 Assume the contrary, i.e., $c = 0$. Then, due to the continuity of $\boldsymbol{\theta} \mapsto \operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])$, there exists a minimizing and convergent sequence in $\mathbb{S}_1$ whose limit $\boldsymbol{\theta}$ is attained and satisfies $\operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) = 0$, i.e.,

$$\int_{\mathbb{R}^2} |\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])|^2 \, \mathrm{d}\mu(\mathbf{x}) = 0.$$

Hence, the support of μ is contained in the line $\{\mathbf{x} \in \mathbb{R}^2 \mid \langle \mathbf{x}, \boldsymbol{\theta} \rangle = \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])\}$ in contradiction to $\mu \in \mathcal{P}_{\mathrm{c}}^*(\mathbb{R}^2)$. □

The NR-CDT is nearly invariant under affine transformations up to bijective remappings of the directions, i.e., up to a resorting of the family $\{\mathcal{N}_{\boldsymbol{\theta}}[\mu] \mid \boldsymbol{\theta} \in \mathbb{S}_1\}$.

Proposition 5 *Let $\mu \in \mathcal{P}_{\mathrm{c}}^*(\mathbb{R}^2)$, $\mathbf{A} \in \mathrm{GL}(2)$, $\mathbf{y} \in \mathbb{R}^2$, and $\mu_{\mathbf{A},\mathbf{y}}$ as in (2). Then, for any $\boldsymbol{\theta} \in \mathbb{S}_1$, the NR-CDT satisfies*

$$\mathcal{N}_{\boldsymbol{\theta}}[\mu_{\mathbf{A},\mathbf{y}}] = \mathcal{N}_{\frac{\mathbf{A}^\top \boldsymbol{\theta}}{\|\mathbf{A}^\top \boldsymbol{\theta}\|}}[\mu].$$

Proof Transferring Proposition 3 to the CDT space, we have

$$\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_{\mathbf{A},\mathbf{y}}](t) = \|\mathbf{A}^\top \boldsymbol{\theta}\| \, \widehat{\mathcal{R}}_{\phi_{\mathbf{A}}(\boldsymbol{\theta})}[\mu](t) + \langle \mathbf{y}, \boldsymbol{\theta} \rangle$$

with the bijection $\phi_{\mathbf{A}}(\boldsymbol{\theta}) \coloneqq (\mathbf{A}^\top \boldsymbol{\theta})/\|\mathbf{A}^\top \boldsymbol{\theta}\|$, $\boldsymbol{\theta} \in \mathbb{S}_1$; so that

$$\operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_{\mathbf{A},\mathbf{y}}]) = \|\mathbf{A}^\top \boldsymbol{\theta}\| \operatorname{mean}(\widehat{\mathcal{R}}_{\phi_{\mathbf{A}}(\boldsymbol{\theta})}[\mu]) + \langle \mathbf{y}, \boldsymbol{\theta} \rangle$$

and

$$\operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu_{\mathbf{A},\mathbf{y}}]) = \|\mathbf{A}^\top \boldsymbol{\theta}\| \operatorname{std}(\widehat{\mathcal{R}}_{\phi_{\mathbf{A}}(\boldsymbol{\theta})}[\mu]).$$

Consequently,

$$\begin{aligned} \mathcal{N}_{\boldsymbol{\theta}}[\mu_{\mathbf{A},\mathbf{y}}](t) &= \frac{\widehat{\mathcal{R}}_{\phi_{\mathbf{A}}(\boldsymbol{\theta})}[\mu](t) - \operatorname{mean}(\widehat{\mathcal{R}}_{\phi_{\mathbf{A}}(\boldsymbol{\theta})}[\mu])}{\operatorname{std}(\widehat{\mathcal{R}}_{\phi_{\mathbf{A}}(\boldsymbol{\theta})}[\mu])} \\ &= \mathcal{N}_{\phi_{\mathbf{A}}(\boldsymbol{\theta})}[\mu](t). \end{aligned}$$

□

3.3 Max-Normalized R-CDT

In the final normalization step, we treat the resorting of $\{\mathcal{N}_{\boldsymbol{\theta}}[\mu] \mid \boldsymbol{\theta} \in \mathbb{S}_1\}$. Since the underlying mapping is unknown in general, we propose to take the supremum over all directions. More precisely, for $\mu \in \mathcal{P}_{\mathrm{c}}^*(\mathbb{R}^2)$, we define its *max-normalized R-CDT* ($_{\mathrm{m}}$NR-CDT) $\mathcal{N}_{\mathrm{m}}[\mu] \colon \mathbb{R} \to \mathbb{R}$ via

$$\mathcal{N}_{\mathrm{m}}[\mu](t) \coloneqq \sup_{\boldsymbol{\theta} \in \mathbb{S}_1} \mathcal{N}_{\boldsymbol{\theta}}[\mu](t), \quad t \in \mathbb{R}.$$

We show that $\mathcal{N}_{\mathrm{m}}$ maps a given measure to a bounded function so that the $_{\mathrm{m}}$NR-CDT space $\mathcal{N}_{\mathrm{m}}[\mathcal{P}_{\mathrm{c}}^*(\mathbb{R}^2)]$ is contained in $L_\rho^\infty(\mathbb{R})$ for the underlying reference measure $\rho \in \mathcal{P}(\mathbb{R})$.

Proposition 6 *Let $\mu \in \mathcal{P}_{\mathrm{c}}^*(\mathbb{R}^2)$. Then, $\mathcal{N}_{\mathrm{m}}[\mu] \in L_\rho^\infty(\mathbb{R})$.*

Proof The restricted Radon operator cannot enlarge the size of the support $\operatorname{diam}(\mu) \coloneqq \sup_{\mathbf{x},\mathbf{y} \in \operatorname{supp}(\mu)} \|\mathbf{x} - \mathbf{y}\|$, i.e., $\operatorname{diam}(\mathcal{R}_{\boldsymbol{\theta}}[\mu]) \leq \operatorname{diam}(\mu)$. Moreover, the range of $\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]$ coincides with the support of $\mathcal{R}_{\boldsymbol{\theta}}[\mu]$. Using that the mean lies in the convex hull of the support, we thus have

$$|\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t) - \operatorname{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])| \leq \operatorname{diam}(\mu) \quad \forall \, \boldsymbol{\theta} \in \mathbb{S}_1.$$

Since $\mu \in \mathcal{P}_{\mathrm{c}}^*(\mathbb{R}^2)$, according to Proposition 4 we have $c \coloneqq \min_{\boldsymbol{\theta} \in \mathbb{S}_1} \operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) > 0$. Thus, the $_{\mathrm{m}}$NR-CDT is bounded by $|\mathcal{N}_{\mathrm{m}}[\mu](t)| \leq \operatorname{diam}(\mu)/c$ for all $t \in \mathbb{R}$. □

With the $_{\mathrm{m}}$NR-CDT, we accomplish our objective to define a transport-based transform that is invariant under affine transformations.

Proposition 7 *Let $\mu \in \mathcal{P}_{\mathrm{c}}^*(\mathbb{R}^2)$, $\mathbf{A} \in \mathrm{GL}(2)$, $\mathbf{y} \in \mathbb{R}^2$, and $\mu_{\mathbf{A},\mathbf{y}}$ as in (2). Then, the $_{\mathrm{m}}$NR-CDT satisfies $\mathcal{N}_{\mathrm{m}}[\mu_{\mathbf{A},\mathbf{y}}] = \mathcal{N}_{\mathrm{m}}[\mu]$.*

Proof Since the mapping $\phi_{\mathbf{A}}(\boldsymbol{\theta}) \coloneqq (\mathbf{A}^\top \boldsymbol{\theta})/\|\mathbf{A}^\top \boldsymbol{\theta}\|$ is a bijection on $\mathbb{S}_1$, we obtain

$$\begin{aligned} \mathcal{N}_{\mathrm{m}}[\mu_{\mathbf{A},\mathbf{y}}](t) &= \sup_{\boldsymbol{\theta} \in \mathbb{S}_1} \mathcal{N}_{\boldsymbol{\theta}}[\mu_{\mathbf{A},\mathbf{y}}](t) \\ &= \sup_{\boldsymbol{\theta} \in \mathbb{S}_1} \mathcal{N}_{\phi_{\mathbf{A}}(\boldsymbol{\theta})}[\mu](t) \\ &= \mathcal{N}_{\mathrm{m}}[\mu](t). \end{aligned}$$

□

The invariance under affine transformations immediately yields the linear separability of affine measure classes, which originate from a single template.

Theorem 1 *For template measures $\mu_0, \nu_0 \in \mathcal{P}_{\mathrm{c}}^*(\mathbb{R}^2)$ with*

$$\mathcal{N}_{\mathrm{m}}[\mu_0] \neq \mathcal{N}_{\mathrm{m}}[\nu_0]$$

consider the classes

$$\mathbb{F} = \{(\mathbf{A} \cdot + \mathbf{y})_{\#}\mu_0 \mid \mathbf{A} \in \mathrm{GL}(2), \, \mathbf{y} \in \mathbb{R}^2\}, \tag{3a}$$

$$\mathbb{G} = \{(\mathbf{A} \cdot + \mathbf{y})_{\#}\nu_0 \mid \mathbf{A} \in \mathrm{GL}(2), \, \mathbf{y} \in \mathbb{R}^2\}. \tag{3b}$$

Then, $\mathbb{F} \subset \mathcal{P}_{\mathrm{c}}^*(\mathbb{R}^2)$ and $\mathbb{G} \subset \mathcal{P}_{\mathrm{c}}^*(\mathbb{R}^2)$ are linearly separable in $_{\mathrm{m}}$NR-CDT space.

Proof Due to the affine construction of $\mathbb{F}$ and $\mathbb{G}$, Proposition 7 yields $\mathcal{N}_\mathrm{m}[\mathbb{F}] = \{\mathcal{N}_\mathrm{m}[\mu_0]\}$ and $\mathcal{N}_\mathrm{m}[\mathbb{G}] = \{\mathcal{N}_\mathrm{m}[\nu_0]\}$. Hence, the assumption $\mathcal{N}_\mathrm{m}[\mu_0] \neq \mathcal{N}_\mathrm{m}[\nu_0]$ implies the linear separability of $\mathcal{N}_\mathrm{m}[\mathbb{F}]$ and $\mathcal{N}_\mathrm{m}[\mathbb{G}]$ in $L^\infty_\rho(\mathbb{R})$. □

The proof of Theorem 1 reveals that the classes $\mathbb{F}$ and $\mathbb{G}$ collapse to single points in ${}_\mathrm{m}$NR-CDT space, directly allowing for linear separability. This is in contrast to the R-CDT results in [23], which can only tackle a subset of affine transforms, and where the considered classes form convex sets in R-CDT space.

3.4 *h*-Normalized R-CDT

We now describe a more general final normalization step than proposed in § 3.3 for treating the resorting of $\{\mathcal{N}_{\boldsymbol{\theta}}[\mu] \mid \boldsymbol{\theta} \in \mathbb{S}_1\}$. To this end, let $h\colon L^\infty(\mathbb{S}_1) \to \mathbb{R}$ be boundedness-preserving, i.e., bounded sets are mapped to bounded sets, such that $h(g \circ \phi_{\mathbf{A}}) = h(g)$ for all $g \in \mathrm{L}^\infty(\mathbb{S}_1)$ and $\phi_{\mathbf{A}}(\boldsymbol{\theta}) = (\mathbf{A}^\top\boldsymbol{\theta})/\|\mathbf{A}^\top\boldsymbol{\theta}\|$ with $\mathbf{A} \in \mathrm{GL}(2)$. With this, for $\mu \in \mathcal{P}^*_\mathrm{c}(\mathbb{R}^2)$, we define its *h-normalized R-CDT* (${}_h$NR-CDT) $\mathcal{N}_h[\mu]\colon \mathbb{R} \to \mathbb{R}$ via

$$\mathcal{N}_h[\mu](t) \coloneqq h(\mathcal{N}[\mu](t,\cdot)), \quad t \in \mathbb{R}.$$

We show that $\mathcal{N}_h$ maps a given measure to a bounded function so that the ${}_h$NR-CDT space $\mathcal{N}_h[\mathcal{P}^*_\mathrm{c}(\mathbb{R}^2)]$ is contained in $L^\infty_\rho(\mathbb{R})$ for the underlying reference measure $\rho \in \mathcal{P}(\mathbb{R})$.

Proposition 8 *$\mathcal{N}_h[\mu] \in L^\infty_\rho(\mathbb{R})$ for all $\mu \in \mathcal{P}^*_\mathrm{c}(\mathbb{R}^2)$.*

Proof As in the proof of Proposition 6, we have

$$|\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t) - \mathrm{mean}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])| \leq \mathrm{diam}(\mu) \quad \forall\, \boldsymbol{\theta} \in \mathbb{S}_1$$

and $c \coloneqq \min_{\boldsymbol{\theta}\in\mathbb{S}_1} \mathrm{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu]) > 0$. Thus, the NR-CDT is uniformly bounded by $|\mathcal{N}[\mu](t,\boldsymbol{\theta})| \leq \mathrm{diam}(\mu)/c$ for all $t \in \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{S}_1$ so that $\mathcal{N}[\mu](t,\cdot) \in L^\infty(\mathbb{S}_1)$. Since $h\colon L^\infty(\mathbb{S}_1) \to \mathbb{R}$ is bounded, there exists a constant $C > 0$ such that $|\mathcal{N}_h[\mu](t)| \leq C$ for all $t \in \mathbb{R}$. □

With the ${}_h$NR-CDT, we have defined a whole family of transport-based transforms that are invariant under affine transformations.

Proposition 9 *Let $\mu \in \mathcal{P}^*_\mathrm{c}(\mathbb{R}^2)$, and $\mu_{\mathbf{A},\mathbf{y}}$ as in (2) with $\mathbf{A} \in \mathrm{GL}(2)$, $\mathbf{y} \in \mathbb{R}^2$. Then, the ${}_h$NR-CDT satisfies $\mathcal{N}_h[\mu_{\mathbf{A},\mathbf{y}}] = \mathcal{N}_h[\mu]$.*

Proof By assumption on h, we directly obtain

$$\begin{aligned}\mathcal{N}_h[\mu_{\mathbf{A},\mathbf{y}}](t) &= h(\mathcal{N}[\mu_{\mathbf{A},\mathbf{y}}](t,\cdot)) \\ &= h(\mathcal{N}[\mu](t,\cdot) \circ \phi_{\mathbf{A}}) \\ &= h(\mathcal{N}[\mu](t,\cdot)) = \mathcal{N}_h[\mu](t)\end{aligned}$$

for all $t \in \mathbb{R}$. □

The invariance under affine transformations immediately yields linear separability of affine measure classes originating from single templates.

Theorem 2 *For template measures $\mu_0, \nu_0 \in \mathcal{P}^*_\mathrm{c}(\mathbb{R}^2)$ with*

$$\mathcal{N}_h[\mu_0] \neq \mathcal{N}_h[\nu_0]$$

consider the classes

$$\mathbb{F} = \big\{(\mathbf{A}\cdot+\mathbf{y})_\#\mu_0 \mid \mathbf{A} \in \mathrm{GL}(2),\, \mathbf{y} \in \mathbb{R}^2\big\}, \tag{4a}$$

$$\mathbb{G} = \big\{(\mathbf{A}\cdot+\mathbf{y})_\#\nu_0 \mid \mathbf{A} \in \mathrm{GL}(2),\, \mathbf{y} \in \mathbb{R}^2\big\}. \tag{4b}$$

Then, $\mathbb{F} \subset \mathcal{P}^*_\mathrm{c}(\mathbb{R}^2)$ and $\mathbb{G} \subset \mathcal{P}^*_\mathrm{c}(\mathbb{R}^2)$ are linearly separable in ${}_h$NR-CDT space.

Proof By construction of $\mathbb{F}$ and $\mathbb{G}$, Proposition 9 yields $\mathcal{N}_h[\mathbb{F}] = \{\mathcal{N}_h[\mu_0]\}$ and $\mathcal{N}_h[\mathbb{G}] = \{\mathcal{N}_h[\nu_0]\}$. Hence, the assumption $\mathcal{N}_h[\mu_0] \neq \mathcal{N}_h[\nu_0]$ implies the linear separability of $\mathcal{N}_h[\mathbb{F}]$ and $\mathcal{N}_h[\mathbb{G}]$ in $L^\infty_\rho(\mathbb{R})$. □

If $\eta\colon \mathbb{R} \to \mathbb{R}$ is a boundedness-preserving function and $H\colon \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ is bounded, where $\mathcal{B}(\mathbb{R})$ denotes the set of bounded subsets of $\mathbb{R}$, we set

$$h(g) = H(\{\eta(g(\boldsymbol{\theta})) \mid \boldsymbol{\theta} \in \mathbb{S}_1\}), \quad g \in L^\infty(\mathbb{S}_1),$$

so that $h(g \circ \phi) = h(g)$ is automatically satisfied for all $g \in \mathrm{L}^\infty(\mathbb{S}_1)$ and all bijections $\phi\colon \mathbb{S}_1 \to \mathbb{S}_1$.

Note that $\eta = \mathrm{Id}$ and $H = \sup$ recovers the ${}_\mathrm{m}$NR-CDT from Section 3.3. The first obvious variation would be choosing $\eta = \mathrm{Id}$ and $H = \inf$. The induced operator, however, does not contain additional information as compared to the ${}_\mathrm{m}$NR-CDT. This is shown in the next proposition, where, for simplicity, we assume that the reference measure ρ is given by the uniform measure $u_{[0,1]}$.

Proposition 10 *Let $\rho = u_{[0,1]}$ and $\mu \in \mathcal{P}^*_\mathrm{c}(\mathbb{R}^2)$. Then,*

$$\inf_{\boldsymbol{\theta}\in\mathbb{S}_1} \mathcal{N}_{\boldsymbol{\theta}}[\mu](t) = -\mathcal{N}_\mathrm{m}[\mu](1-t) \quad \forall\, t \in \mathbb{R}.$$

Proof As $\rho = u_{[0,1]}$ by assumption, for all $t \in \mathbb{R}$, we have

$$F_{u_{[0,1]}}(1-t) = 1 - F_{u_{[0,1]}}(t).$$

Moreover, for $\tau \in [0,1]$,

$$F^{[-1]}_{\mathcal{R}_{\boldsymbol{\theta}}[\mu]}(1-\tau) = -F^{[-1]}_{\mathcal{R}_{-\boldsymbol{\theta}}[\mu]}(\tau),$$

since the Radon transform satisfies the eveness property

$$\mathcal{R}_{\boldsymbol{\theta}}[\mu]((s,\infty)) = \mathcal{R}_{-\boldsymbol{\theta}}[\mu]((-\infty,-s)) \quad \forall\, s \in \mathbb{R}.$$

This implies that

$$\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](1-t) = -\widehat{\mathcal{R}}_{-\boldsymbol{\theta}}[\mu](t)$$

and, consequently,

$$\begin{aligned}\inf_{\boldsymbol{\theta}\in\mathbb{S}_1} \mathcal{N}_{\boldsymbol{\theta}}[\mu](t) &= -\sup_{\boldsymbol{\theta}\in\mathbb{S}_1} \frac{-\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu](t) - \operatorname{mean}(-\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])}{\operatorname{std}(-\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])} \\ &= -\sup_{\boldsymbol{\theta}\in\mathbb{S}_1} \frac{\widehat{\mathcal{R}}_{-\boldsymbol{\theta}}[\mu](1-t) - \operatorname{mean}(\widehat{\mathcal{R}}_{-\boldsymbol{\theta}}[\mu])}{\operatorname{std}(\widehat{\mathcal{R}}_{-\boldsymbol{\theta}}[\mu])} \\ &= -\mathcal{N}_{\mathrm{m}}[\mu](1-t),\end{aligned}$$

as stated. □

The following example lists further variations of the h-normalization based on η and H.

Example 1 Suitable choices of η and H include

a) $\eta = |\cdot|$ and $H = \sup$, i.e.,
$$\mathcal{N}_{h_{\mathrm{a}}}[\mu](t) = \sup_{\boldsymbol{\theta}\in\mathbb{S}_1} |\mathcal{N}_{\boldsymbol{\theta}}[\mu](t)|,$$

b) $\eta = |\cdot|$ and $H = \inf$, i.e.,
$$\mathcal{N}_{h_{\mathrm{b}}}[\mu](t) = \inf_{\boldsymbol{\theta}\in\mathbb{S}_1} |\mathcal{N}_{\boldsymbol{\theta}}[\mu](t)|,$$

c) $\eta = |\cdot|$ and $H = \sup - \inf$, i.e.,
$$\mathcal{N}_{h_{\mathrm{c}}}[\mu](t) = \sup_{\boldsymbol{\theta}\in\mathbb{S}_1} |\mathcal{N}_{\boldsymbol{\theta}}[\mu](t)| - \inf_{\boldsymbol{\theta}\in\mathbb{S}_1} |\mathcal{N}_{\boldsymbol{\theta}}[\mu](t)|.$$

d) $\eta = \mathrm{Id}$ and $H = \sup - \inf$, i.e.,
$$\mathcal{N}_{h_{\mathrm{d}}}[\mu](t) = \sup_{\boldsymbol{\theta}\in\mathbb{S}_1} \mathcal{N}_{\boldsymbol{\theta}}[\mu](t) - \inf_{\boldsymbol{\theta}\in\mathbb{S}_1} \mathcal{N}_{\boldsymbol{\theta}}[\mu](t),$$

Note that $\mathcal{N}_{h_{\mathrm{a}}}[\mu]$ and $\mathcal{N}_{h_{\mathrm{b}}}[\mu]$ capture only a single angular piece of information, the maximum or minimum, respectively, whereas $\mathcal{N}_{h_{\mathrm{c}}}[\mu]$ and $\mathcal{N}_{h_{\mathrm{d}}}[\mu]$ encode the angular range. To include even further knowledge, we now construct an example for a normalization not only operating on the image $\{\mathcal{N}_{\boldsymbol{\theta}}[\mu](t) \mid \boldsymbol{\theta} \in \mathbb{S}_1\}$. To this end, we introduce the *TV-normalized NR-CDT* ($_{\mathrm{tv}}$NR-CDT) $\mathcal{N}_{\mathrm{tv}}[\mu]\colon \mathbb{R} \to \mathbb{R}$ via

$$\mathcal{N}_{\mathrm{tv}}[\mu](t) \coloneqq \sup_{P\in\mathbb{P}} \sum_{i=1}^{n_P} |\mathcal{N}_{\boldsymbol{\theta}_{i+1}}[\mu](t) - \mathcal{N}_{\boldsymbol{\theta}_i}[\mu](t)|,$$

where the supremum runs over the set of partitions $\mathbb{P} = \{P = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n_P+1}) \mid P \text{ is partition of } \mathbb{S}_1\}$, which means that there exist $0 \le \vartheta_1 < \vartheta_{n_P} < 2\pi$ so that $\boldsymbol{\theta}_i = (\cos(\vartheta_i), \sin(\vartheta_i))^\top$ and $\boldsymbol{\theta}_{n_P+1} = \boldsymbol{\theta}_1$. To ensure well-definedness, we restrict ourselves to the class

$$\mathcal{P}^*_{\mathrm{tv}}(\mathbb{R}^2) \coloneqq \{\mu \in \mathcal{P}^*_{\mathrm{c}}(\mathbb{R}^2) \mid \mathcal{N}_{\mathrm{tv}}[\mu](t) < \infty \,\forall\, t \in \mathbb{R}\}, \tag{5}$$

which is a suitable setting for all our numerical experiments below. This normalization variant corresponds to

$$h(g) = \sup_{P\in\mathbb{P}} \sum_{i=1}^{n_P} |g(\boldsymbol{\theta}_{i+1}) - g(\boldsymbol{\theta}_i)|,$$

which satisfies $h(g \circ \phi_{\mathbf{A}}) = h(g)$ for all $\phi_{\mathbf{A}}(\boldsymbol{\theta}) = (\mathbf{A}^\top\boldsymbol{\theta})/\|\mathbf{A}^\top\boldsymbol{\theta}\|$ with $\mathbf{A} \in \mathrm{GL}(2)$ since $\phi_A(\mathbb{P}) = \mathbb{P}$. Therefore, Theorem 2 also holds for $_{\mathrm{tv}}$NR-CDT, when assuming $\mu_0, \nu_0 \in \mathcal{P}^*_{\mathrm{tv}}(\mathbb{R}^2)$.

4 Generalized NR-CDT

The model behind $_h$NR-CDT is clearly tailored to 2d pattern recognition tasks under affine transformations. Theoretically, this idea may be transferred to the multi-dimensional and non-Euclidean setting. For the underlying R-CDT and absolutely continuous probabilities, this extension is studied in [11], where the Radon transform is replaced by the so-called generalized Radon transform.

To extend this approach beyond functions, and to allow more flexibility, we consider arbitrary (probability) measures on a Polish space $\mathbb{X}$, i.e., a separable completely metrizable topological space. On the basis of a *direction set* Θ and a *defining function* $\phi\colon \mathbb{X}\times\Theta \to \mathbb{R}$ such that $\phi(\cdot,\boldsymbol{\theta})\colon \mathbb{X} \to \mathbb{R}$ is measurable for all $\boldsymbol{\theta} \in \Theta$, we define the *generalized slicing operator* $S_{\phi,\boldsymbol{\theta}}\colon \mathbb{X} \to \mathbb{R}$ by

$$S_{\phi,\boldsymbol{\theta}}(\mathbf{x}) \coloneqq \phi(\mathbf{x},\boldsymbol{\theta}), \quad \mathbf{x} \in \mathbb{X},\ \boldsymbol{\theta} \in \Theta, \tag{6}$$

and the *generalized restricted Radon transform* via

$$\mathcal{R}_{\phi,\boldsymbol{\theta}}\colon \mathcal{M}(\mathbb{X}) \to \mathcal{M}(\mathbb{R}), \quad \mu \mapsto (S_{\phi,\boldsymbol{\theta}})_{\#}\mu. \tag{7}$$

If Θ is a Polish space too, then the restricted Radon transforms may be glued together to form a measure on $\mathbb{R}\times\Theta$. More precisely, for a *gluing measure* $\gamma \in \mathcal{P}(\Theta)$, we define the *generalized*

Radon transform $\mathcal{R}_{\phi,\gamma}\colon \mathcal{M}(\mathbb{X}) \to \mathcal{M}(\mathbb{R}\times\Theta)$ as

$$\mathcal{R}_{\phi,\gamma}[\mu] \coloneqq (\mathcal{I}_\phi)_\#[\mu\times\gamma]$$

with $\mathcal{I}_\phi(\mathbf{x},\boldsymbol{\theta}) \coloneqq (S_{\phi,\boldsymbol{\theta}}(\mathbf{x}),\boldsymbol{\theta})$ for $(\mathbf{x},\boldsymbol{\theta})\in\mathbb{X}\times\Theta$. Similarly to before, the generalized (restricted) Radon transform maps probability measures to probability measures. Furthermore, the disintegration of the Radon transform carries over.

Proposition 11 *Let $\mu\in\mathcal{M}(\mathbb{X})$. Then, $\mathcal{R}_{\phi,\gamma}[\mu]$ can be disintegrated into the family $\mathcal{R}_{\phi,\boldsymbol{\theta}}[\mu]$ with respect to γ, i.e., for all continuous $g\in C_0(\mathbb{R}\times\Theta)$ vanishing at infinity, we have*

$$\langle \mathcal{R}_{\phi,\gamma}[\mu], g\rangle = \int_\Theta \langle \mathcal{R}_{\phi,\boldsymbol{\theta}}[\mu], g(\cdot,\boldsymbol{\theta})\rangle \,\mathrm{d}\gamma(\boldsymbol{\theta}).$$

Proof Using Fubini's theorem, we directly obtain

$$\begin{aligned}\langle \mathcal{R}_{\phi,\gamma}[\mu], g\rangle &= \int_{\mathbb{R}\times\Theta} g(t,\boldsymbol{\theta})\,\mathrm{d}\mathcal{I}_\#[\mu\times\gamma](t,\boldsymbol{\theta})\\ &= \int_\Theta\int_{\mathbb{X}} g(S_{\phi,\boldsymbol{\theta}}(\mathbf{x}),\boldsymbol{\theta})\,\mathrm{d}\mu(\mathbf{x})\,\mathrm{d}\gamma(\boldsymbol{\theta})\\ &= \int_\Theta\int_{\mathbb{X}} g(t,\boldsymbol{\theta})\,\mathrm{d}[(S_{\phi,\boldsymbol{\theta}})_\#\mu](t)\,\mathrm{d}\gamma(\boldsymbol{\theta})\\ &= \int_\Theta \langle \mathcal{R}_{\phi,\boldsymbol{\theta}}[\mu], g(\cdot,\boldsymbol{\theta})\rangle\,\mathrm{d}\gamma(\boldsymbol{\theta}).\end{aligned}$$ □

Notice that the original generalized Radon transform for smooth functions is based on a so-called double fibering [34]. If ϕ satisfies certain regularity assumptions, this generalized Radon transform becomes invertible [34–37]. Our generalized Radon transform for measures is closely related to the *back projection*

$$\mathcal{R}^*_{\phi,\gamma}[g](\mathbf{x}) \coloneqq \int_\Theta g(S_{\phi,\boldsymbol{\theta}}(\mathbf{x}),\theta)\,\mathrm{d}\gamma(\boldsymbol{\theta}),\quad \mathbf{x}\in\mathbb{X}.$$

Proposition 12 *The generalized Radon transform of $\mu\in\mathcal{M}(\mathbb{X})$ satisfies*

$$\langle \mathcal{R}_{\phi,\gamma}[\mu], g\rangle = \langle \mu, \mathcal{R}^*_{\phi,\gamma}[g]\rangle \quad \forall\, g\in L^\infty_{\lambda_\mathbb{R}\times\gamma}(\mathbb{R}\times\Theta).$$

Proof For all $\mu\in\mathcal{M}(\mathbb{X})$ and $g\in L^\infty(\mathbb{R}\times\Theta)$, applying Fubini's theorem gives

$$\begin{aligned}\langle \mathcal{R}_{\phi,\gamma}[\mu], g\rangle &= \int_{\mathbb{R}\times\Theta} g(t,\boldsymbol{\theta})\,\mathrm{d}(\mathcal{I}_\phi)_\#[\mu\times\gamma](t,\boldsymbol{\theta})\\ &= \int_{\mathbb{X}}\int_\Theta g(S_{\phi,\boldsymbol{\theta}}(\mathbf{x}),\boldsymbol{\theta})\,\mathrm{d}\gamma(\boldsymbol{\theta})\,\mathrm{d}\mu(\mathbf{x}).\end{aligned}$$ □

To introduce a generalized ${}_h$NR-CDT, we have to ensure that each normalization step is well-defined. For this, we restrict ourselves to *uniformly bounded* defining functions ϕ meaning that, for every $L>0$, there exits $M>0$ satisfying

$$|\phi(\mathbf{x},\boldsymbol{\theta})|\le M \qquad \forall\,\|\mathbf{x}\|\le L,\ \forall\,\boldsymbol{\theta}\in\Theta.$$

Furthermore, we only consider measures from

$$\begin{aligned}\mathcal{P}^+_{\phi,\mathrm{c}}(\mathbb{X}) \coloneqq \{\mu\in\mathcal{P}(\mathbb{X}) \mid \operatorname{supp}(\mu)\subset\subset\mathbb{X}\wedge\\ \exists\, c>0 : \operatorname{std}(\mathcal{R}_{\phi,\boldsymbol{\theta}}[\mu])\ge c\ \forall\,\theta\in\Theta\}.\end{aligned}$$

For uniformly bounded ϕ and $\mu\in\mathcal{P}^+_\mathrm{c}(\mathbb{X})$, the support of $\mathcal{R}_{\phi,\boldsymbol{\theta}}[\mu]$ is bounded; so their means and standard derivations exist.

The *generalized NR-CDT* $\mathcal{N}_\phi[\mu]\colon\mathbb{R}\times\Theta\to\mathbb{R}$ of $\mu\in\mathcal{P}^+_{\phi,\mathrm{c}}(\mathbb{X})$ is defined as

$$\mathcal{N}_\phi[\mu](t,\boldsymbol{\theta}) \coloneqq \mathcal{N}_{\phi,\boldsymbol{\theta}}[\mu](t)$$

with

$$\mathcal{N}_{\phi,\boldsymbol{\theta}}[\mu](t) \coloneqq \frac{\widehat{\mathcal{R}}_{\phi,\boldsymbol{\theta}}[\mu](t) - \operatorname{mean}(\widehat{\mathcal{R}}_{\phi,\boldsymbol{\theta}}[\mu])}{\operatorname{std}(\widehat{\mathcal{R}}_{\phi,\boldsymbol{\theta}}[\mu])}$$

By construction, $\mathcal{N}_\phi[\mu]$ is bounded on $\operatorname{supp}(\rho)\times\Theta$. In the style of § 3.4, let $h\colon L^\infty(\Theta)\to\mathbb{R}$ be bounded. Unless otherwise stated, we assume that $h(g\circ\psi)=h(g)$ for all $g\in L^\infty(\Theta)$ and any bijection $\psi\colon\Theta\to\Theta$, which is a stricter setting as in § 3.4 to allow for a general discussion. For $\mu\in\mathcal{P}^+_{\phi,\mathrm{c}}(\mathbb{X})$, the *generalized ${}_h$NR-CDT* $\mathcal{N}_{h,\phi}[\mu]\colon\mathbb{R}\to\mathbb{R}$ may be defined via

$$\mathcal{N}_{h,\phi}[\mu](t) \coloneqq h(\mathcal{N}_\phi[\mu](t,\cdot)),\quad t\in\mathbb{R}.$$

In the following, we briefly discuss the situation for specific $\mathbb{X}$ and ϕ, especially with respect to the possible invariant transformations.

4.1 Multi-dimensional NR-CDT

Choosing $\mathbb{X}\coloneqq\mathbb{R}^d$ and $\Theta\coloneqq\mathbb{S}_{d-1}\coloneqq\{\mathbf{x}\in\mathbb{R}^d \mid |\mathbf{x}|=1\}$ together with Euclidean inner product

$$\phi(\mathbf{x},\boldsymbol{\theta}) \coloneqq \langle\mathbf{x},\boldsymbol{\theta}\rangle,\quad \mathbf{x}\in\mathbb{R}^d,\boldsymbol{\theta}\in\mathbb{S}_{d-1},$$

we obtain the straightforward generalization of our NR-CDT variants to the multi-dimensional setting. A brief inspection of the two-dimensional

setting shows that all results line-by-line generalize to the multi-dimensional setting. In particular, for measures in the class

$$\mathcal{P}_{\mathrm{c}}^{*}(\mathbb{R}^d) \coloneqq \{\mu \in \mathcal{P}(\mathbb{R}^d) \mid \operatorname{supp}(\mu) \subset\subset \mathbb{R}^d \wedge \dim(\operatorname{supp}(\mu)) > d-1\},$$

whose supports are not concentrated on hyperplanes, the standard deviation $\operatorname{std}(\widehat{\mathcal{R}}_{\boldsymbol{\theta}}[\mu])$ is uniformly bounded away from zero, i.e., $\mathcal{P}_{\mathrm{c}}^{*}(\mathbb{R}^d) = \mathcal{P}_{\phi,\mathrm{c}}^{+}(\mathbb{R}^d)$. The multi-dimensional ${}_h$NR-CDT again promotes the linear separability of affine classes.

Theorem 3 *For template measures $\mu_0, \nu_0 \in \mathcal{P}_{\mathrm{c}}^{*}(\mathbb{R}^d)$ with*

$$\mathcal{N}_h[\mu_0] \neq \mathcal{N}_h[\nu_0]$$

consider the classes

$$\mathbb{F} = \left\{(\mathbf{A}\cdot+\mathbf{y})_{\#}\mu_0 \mid \mathbf{A} \in \mathrm{GL}(d),\, \mathbf{y} \in \mathbb{R}^d\right\},$$
$$\mathbb{G} = \left\{(\mathbf{A}\cdot+\mathbf{y})_{\#}\nu_0 \mid \mathbf{A} \in \mathrm{GL}(d),\, \mathbf{y} \in \mathbb{R}^d\right\}.$$

Then, $\mathbb{F} \subset \mathcal{P}_{\mathrm{c}}^{*}(\mathbb{R}^d)$ and $\mathbb{G} \subset \mathcal{P}_{\mathrm{c}}^{*}(\mathbb{R}^d)$ are linearly separable in ${}_h$NR-CDT space.

4.2 Circular NR-CDT

Another example for a generalized Radon transform on the multi-dimensional Euclidean space is the circular Radon transform [38]. Domain and direction set of this transform are given by $\mathbb{X} \coloneqq \mathbb{R}^d$ and $\Theta \coloneqq \mathbb{R}^d$. Furthermore, the slicing operator reads

$$\phi(\mathbf{x}, \boldsymbol{\theta}) \coloneqq \|\mathbf{x} - \boldsymbol{\theta}\|, \quad \mathbf{x}, \boldsymbol{\theta} \in \mathbb{R}^d. \tag{8}$$

Figuratively, the circular Radon transform integrates along circles with center $\boldsymbol{\theta} \in \Theta$. The resulting ${}_h$NR-CDT is especially useful for classification tasks under isotropic affine transformations, i.e., for probability classes build by

$$\mu_{s\mathbf{Q},\mathbf{y}} \coloneqq (s\mathbf{Q}\cdot+\mathbf{y})_{\#}\mu$$

with $s > 0$, $\mathbf{Q} \in \mathrm{O}(d)$ in the orthogonal group, and $\mathbf{y} \in \mathbb{R}^d$.

Proposition 13 *For any $\boldsymbol{\theta} \in \mathbb{R}^d$, the restricted circular Radon transform satisfies*

$$\mathcal{R}_{\phi,\boldsymbol{\theta}}[\mu_{s\mathbf{Q},\mathbf{y}}] = (s\cdot)_{\#}\mathcal{R}_{\phi,\mathbf{Q}^\top(\boldsymbol{\theta}-\mathbf{y})}[\mu].$$

Proof The circular Radon transform is induced by the slicing operator in (8), which here yields

$$\begin{aligned}\mathcal{R}_{\phi,\boldsymbol{\theta}}[\mu_{s\mathbf{Q},\mathbf{y}}] &= (S_{\phi,\boldsymbol{\theta}})_{\#}[(s\mathbf{Q}\cdot+\mathbf{y})_{\#}\mu] \\ &= \|s\mathbf{Q}\cdot+\mathbf{y}-\boldsymbol{\theta}\|_{\#}\mu \\ &= \|s\cdot-\mathbf{Q}^\top(\boldsymbol{\theta}-\mathbf{y})\|_{\#}\mu \\ &= (s\cdot)_{\#}\mathcal{R}_{\phi,\mathbf{Q}^\top(\boldsymbol{\theta}-\mathbf{y})}[\mu].\end{aligned}$$

□

Due to the normalizations behind ${}_h$NR-CDT, Proposition 13 immediately implies

$$\mathcal{N}_{\phi,h}[\mu_{s\mathbf{Q},\mathbf{y}}] = \mathcal{N}_{\phi,h}[h] \tag{9}$$

for all $\mu \in \mathcal{P}_{\phi,\mathrm{c}}^{+}(\mathbb{R}^d)$, $s > 0$, $\mathbf{Q} \in \mathrm{O}(d)$, and $\mathbf{y} \in \mathbb{R}^d$. Therefore, our circular ${}_h$NR-CDT is invariant under isotropic affine transformations and promotes separability of the corresponding classes.

Theorem 4 *For template measures $\mu_0, \nu_0 \in \mathcal{P}_{\phi,\mathrm{c}}^{+}(\mathbb{R}^d)$ whose circular ${}_h$NR-CDTs satisfy*

$$\mathcal{N}_{\phi,h}[\mu_0] \neq \mathcal{N}_{\phi,h}[\nu_0],$$

consider the classes

$$\mathbb{F} = \left\{(s\mathbf{Q}\cdot+\mathbf{y})_{\#}\mu_0 \mid s > 0,\, \mathbf{Q} \in \mathrm{O}(d),\, \mathbf{y} \in \mathbb{R}^d\right\},$$
$$\mathbb{G} = \left\{(s\mathbf{Q}\cdot+\mathbf{y})_{\#}\nu_0 \mid s > 0,\, \mathbf{Q} \in \mathrm{O}(d),\, \mathbf{y} \in \mathbb{R}^d\right\}.$$

Then, $\mathbb{F} \subset \mathcal{P}_{\phi,\mathrm{c}}^{+}(\mathbb{R}^d)$ and $\mathbb{G} \subset \mathcal{P}_{\phi,\mathrm{c}}^{+}(\mathbb{R}^d)$ are linearly separable in circular ${}_h$NR-CDT space.

Proof Because of (9), we have $\mathcal{N}_{\phi,h}[\mathbb{F}] = \{\mathcal{N}_{\phi,h}[\mu_0]\}$ and $\mathcal{N}_{\phi,h}[\mathbb{G}] = \{\mathcal{N}_{\phi,h}[\nu_0]\}$. Then $\mathcal{N}_{\phi,h}[\mu_0] \neq \mathcal{N}_{\phi,h}[\nu_0]$ implies their linear separability in $L_{\rho}^{\infty}(\mathbb{R})$. □

4.3 NR-CDT on the Rotation Group

The generalized ${}_h$NR-CDT is not restricted to the Euclidean setting. As an instance, we consider the Radon transform on the 3d rotation group introduced in [30], which corresponds to $\mathbb{X}, \Theta \coloneqq \mathrm{SO}(3)$ and the slicing operator

$$\phi(\mathbf{x}, \boldsymbol{\theta}) \coloneqq \arccos\left(\frac{\operatorname{trace}(\boldsymbol{\theta}^\top\mathbf{x}) - 1}{2}\right).$$

In the following, we consider the rotation of measures given by $\mu_{\mathbf{R}} \coloneqq (\mathbf{R}\cdot)_{\#}\mu$ for $\mu \in \mathcal{P}_{\phi,\mathrm{c}}^{+}(\mathrm{SO}(3))$ and $\mathbf{R} \in \mathrm{SO}(3)$.

Proposition 14 *For any $\boldsymbol{\theta} \in \mathrm{SO}(3)$, the restricted Radon transform on $\mathrm{SO}(3)$ satisfies*

$$\mathcal{R}_{\phi,\boldsymbol{\theta}}[\mu_{\mathbf{R}}] = \mathcal{R}_{\phi,\mathbf{R}^\top\boldsymbol{\theta}}[\mu].$$

Proof Plugging in the definition of the generalized restricted Radon transform, we obtain

$$\begin{aligned}\mathcal{R}_{\phi,\boldsymbol{\theta}}[\mu_{\mathbf{R}}] &= (S_{\phi,\boldsymbol{\theta}})_{\#}[(\mathbf{R}\,\cdot)_{\#}\mu] \\ &= \Big(\arccos\Big(\tfrac{\operatorname{trace}(\boldsymbol{\theta}^\top\mathbf{R}\,\cdot)-1}{2}\Big)\Big)_{\#}\mu \\ &= \mathcal{R}_{\phi,\mathbf{R}^\top\boldsymbol{\theta}}[\mu]. \qquad \square\end{aligned}$$

Since h has to be invariant under the rotation of the direction set $\Theta = \mathrm{SO}(3)$, the ${}_h$NR-CDT inherit this property, promoting the separability under rotations.

Theorem 5 *For $\mu_0, \nu_0 \in \mathcal{P}^+_{\phi,\mathrm{c}}(\mathrm{SO}(3))$ with*

$$\mathcal{N}_h[\mu_0] \neq \mathcal{N}_h[\nu_0],$$

consider the classes

$$\mathbb{F} = \big\{(\mathbf{R}\cdot)_{\#}\mu_0 \mid \mathbf{R} \in \mathrm{SO}(3)\big\},$$
$$\mathbb{G} = \big\{(\mathbf{R}\cdot)_{\#}\nu_0 \mid \mathbf{R} \in \mathrm{SO}(3)\big\}.$$

Then, $\mathbb{F} \subset \mathcal{P}^+_{\phi,\mathrm{c}}(\mathbb{R}^d)$ and $\mathbb{G} \subset \mathcal{P}^+_{\phi,\mathrm{c}}(\mathbb{R}^d)$ are linearly separable in ${}_h$NR-CDT space.

The proof is literally the same as of Theorem 4. Notice that the Radon transform on SO(3) of a rotated measure causes only a reparametrization of the direction set; therefore, the first normalization step of the ${}_h$NR-CDT is not necessary in the strict sense.

4.4 Further NR-CDT Variants

In addition to the above examples, the ${}_h$NR-CDT can be adapted to further generalized Radon transforms. E.g. for studying measures on spheres, the vertically sliced transform [29] and parallelly sliced transform [30] may be of interest, where our normalization procedure yields invariant feature extractor under rotations. For measures on hyperbolic spaces, the hyperbolic Radon transform [39] may be used. Finally, for measures on arbitrary manifolds, the eigenfunctions of the Laplace–Beltrami operator may be employed to define generalized Radon transforms regarding our framework in the style of the intrinsic sliced Wasserstein distance [40]. Actual studies about the usefulness of the generalized ${}_h$NR-CDT to obtain invariant feature extractors under certain transformations are left for future research.

5 Numerical Experiments

We now present numerical experiments to support our linear separability results in Theorems 1–5. Keeping the application in computational filigranology in mind, we start with the two-dimensional setting and focus on image classification and clustering to show the effectiveness of our new feature representations. Thereon, we provide proof-of-concept experiments in the multi-dimensional setting, where we focus on classification in $\mathbb{R}^3$ and SO(3). We compare our approach with the original R-CDT [23] and the Euclidean representation as baseline. Since R-CDT already outperforms other state-of-the-art approaches in the small data regime [28], we omit comparisons with neural networks. All methods are implemented in Julia and the code is publicly available at https://github.com/DrBeckmann/NR-CDT. Our experiments are performed on a off-the-shelf MacBookPro 2020 with 1.4 GHz quad-core Intel Core i5 CPU and 8 GB RAM.

Before reporting our numerical results, we wish to comment on some details regarding the numerical implementation. In the two-dimensional setting, each image is modelled as a piecewise constant function with support in $[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]^2 \subset \mathbb{B}_2$, which allows us to compute its Radon transform in closed form. For numerical stability, we replace the line integrals in (1) by integrals over disjoint stripes of width $\frac{2}{R}$, where R is the number of radial samples. Thereon, the CDTs are computed based on linear interpolation on an equispaced grid of interpolation points in $(0, 1)$. In the multi-dimensional setting, we model the objects as empirical measures on $\mathbb{X}$. Then, for a discretized direction set Θ, the restricted Radon transforms in (7) can be explicitly computed by evaluating (6), leading to piecewise constant functions on $[0, 1]$. Then, the CDTs are exactly calculated by explicit inversion.

5.1 Image Classification

Our two-dimensional numerical experiments make use of the following three datasets:

a) **Academic dataset.** Based on (up to) three synthetic template symbols, cf. Figure 3 (top row), represented by 256×256 pixels, we

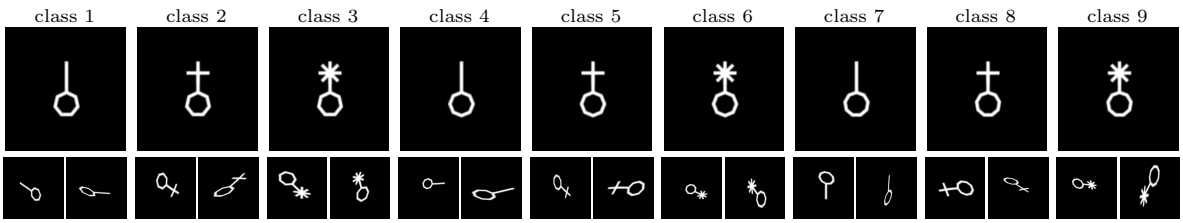

Fig. 2 Synthetic templates of the polygon dataset (top) and random affinely transformed samples including slight non-affine perturbations of the templates (bottom).

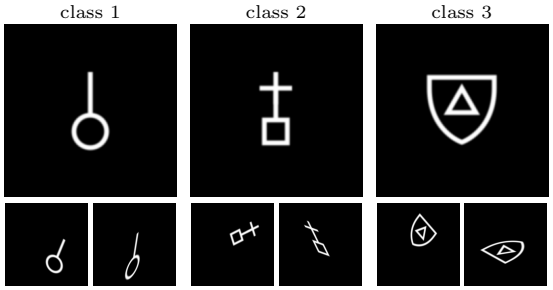

Fig. 3 Synthetic templates of the academic dataset (top) and random affinely transformed samples (bottom).

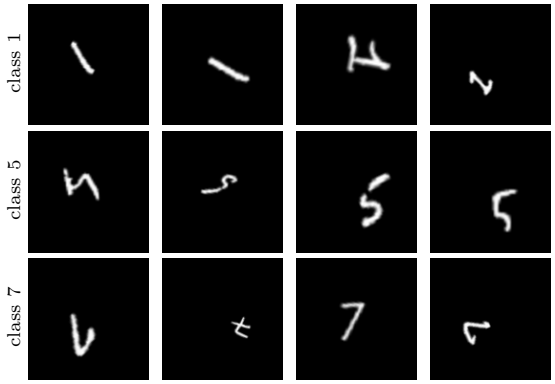

Fig. 4 Samples of the LinMNIST dataset (ones, fives and sevens) based on affinely transformed MNIST digits.

construct our classes by affinely transforming the templates with random pixel shifts in $[-40, 40]$, rotation angles in $[0°, 360°)$, anisotropic scaling in $[0.75, 1.0]$, and shearing in $[-30°, 30°]$, cf. Figure 3 (bottom row). Note that this dataset was first introduced in [32] for proof-of-concept experiments showing linear separability in ${}_{\mathrm{m}}$NR-CDT space.

b) **Polygon dataset.** Nine synthetic template symbols, cf. Figure 2 (top row), each of size 256×256 pixels, are slightly non-affinely deformed and, thereon, affinely transformed with random pixel shifts in $[-20, 20]$, rotation angles in $[0°, 360°)$, anisotropic scaling in $[0.5, 1.25]$, and shearing in $[-30°, 30°]$, cf. Figure 2 (bottom row). For the non-affine deformations, we assign to the (j, k)th pixel the bi-quadratically interpolated gray values at the perturbed location

$$\big(j + a_1 \sin\big(\tfrac{2\pi f_1}{256}\, k\big), k + a_2 \cos\big(\tfrac{2\pi f_2}{256}\, j\big)\big)$$

with random frequencies f_1, f_2 in $[1.5, 2.5]$ and amplitudes a_1, a_2 in $[0.5, 2.0]$.

c) **LinMNIST dataset.** This dataset was first introduced in [41] and consists of affinely transformed MNIST digits [42]. Here, we restrict ourselves to the three classes $\{1, 5, 7\}$, rescale the data to 128×128 pixels and use random pixel shifts in $[-20, 20]$, rotation angles in $[0°, 360°)$, and anisotropic scaling in $[0.5, 1.25]$, see Figure 4 for a random selection of samples of the generated dataset.

5.1.1 Nearest Template Classification

In the following first experiments, we aim to validate the theoretical results from Theorems 1 and 2. To this end, we focus on the first two datasets, which allow us to fully control the occurring affine and non-affine deformations. Looking at the proofs, we observe that the ${}_{h}$NR-CDTs map each entire affine class to a single point in the corresponding feature space. As the datasets originate from known templates, the easiest way for classification is the nearest template (NT) method, which assigns the label of the closest template in the considered feature spaces.

In order to observe the predicted behaviour numerically, the underlying Radon transform and CDT have to be discretized fine enough. Therefor, we choose different numbers of equispaced angles and 850 equispaced radii for the Radon transform

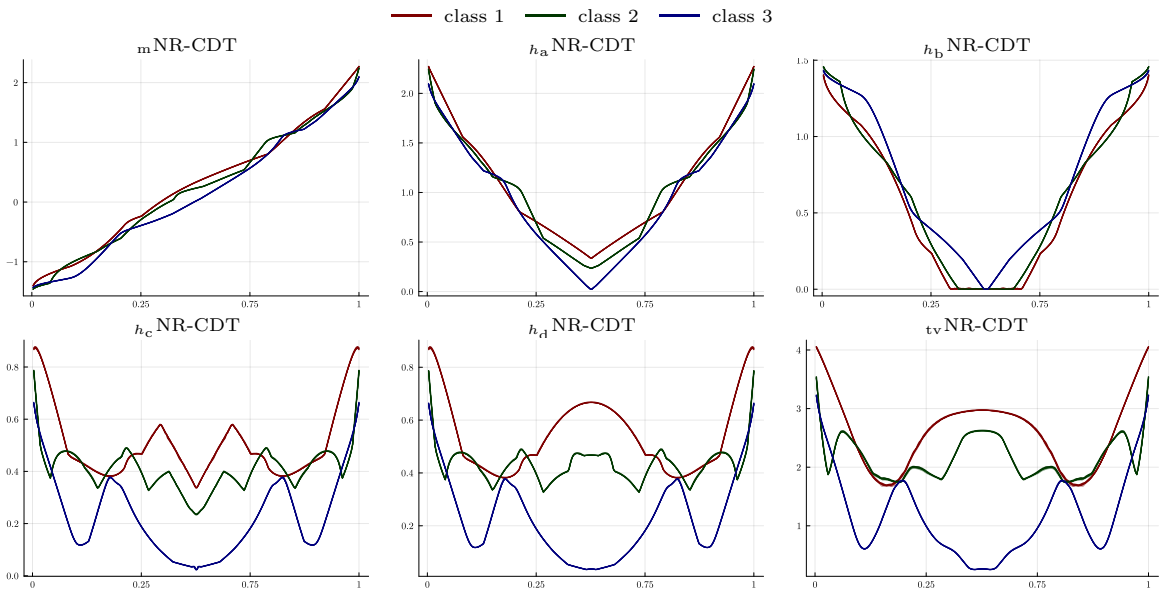

Fig. 5 Visualizations of various $_h$NR-CDTs for the academic dataset with 10 samples per class and 256 Radon angles.

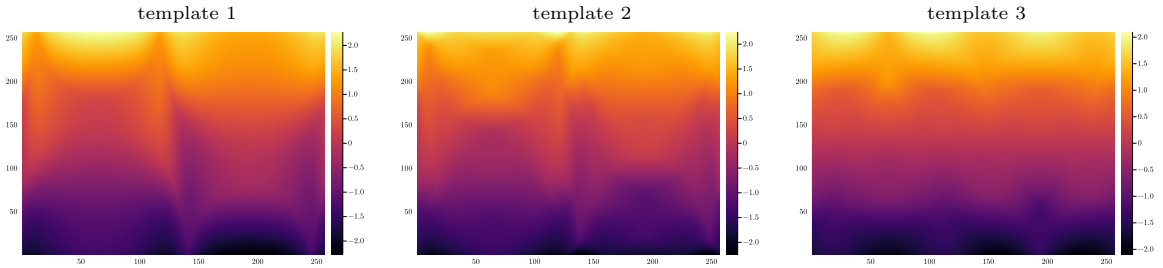

Fig. 6 Visualization of the NR-CDTs for the three template measures of the academic dataset with 256 Radon angles.

as well as 256 equispaced interpolation points for the CDT.

We start with reporting our NT results for the academic dataset with 10 samples per class, where we compare the Chebychev norm $\|\cdot\|_\infty$ and Euclidean norm $\|\cdot\|_2$ in $_h$NR-CDT space for determining the nearest template. For illustration, the $_h$NR-CDTs of the entire dataset are depicted in Figure 5, where the thick lines correspond to the templates and the thin lines to the transformed images. As predicted by our theory, the lines per class are nearly perfectly aligned. This also holds for the $_{\mathrm{tv}}$NR-CDT, which in theory requires more regularity of the measures. An inspection of Figure 6, however, reveals that the corresponding NR-CDTs of the templates are rather smooth.

Our classification results are shown in Table 2 for varying numbers of equispaced angles for the underlying Radon transform. We observe that the $_h$NR-CDT feature representations clearly outperform the R-CDT and Euclidean baseline and, remarkably, the classification is already perfect for a small number of chosen angles. As the different $_h$NR-CDT representations perform nearly on par, we henceforth consider only $_{\mathrm{m}}$NR-CDT, $_{h_{\mathrm{d}}}$NR-CDT and $_{\mathrm{tv}}$NR-CDT. Moreover, in most cases, $\|\cdot\|_2$ outperforms $\|\cdot\|_\infty$ so that from now on we restrict ourselves to the Euclidean norm $\|\cdot\|_2$.

To study the robustness of the NT classification we consider the polygon dataset with 10 samples per class, which includes small non-affine perturbations. Hence, a perfect classification according to Theorems 1 and 2 cannot be expected. The accuracy results are reported in Table 3. While the Euclidean baseline and R-CDT are as bad as random guessing, $_{\mathrm{m}}$NR-CDT and $_{h_{\mathrm{d}}}$NR-CDT perform significantly better and, for sufficiently many Radon angles, $_{\mathrm{tv}}$NR-CDT even

angles	Eucl.		R-CDT		$_{\mathrm{m}}$NR-CDT		$_{h_{\mathrm{a}}}$NR-CDT		$_{h_{\mathrm{b}}}$NR-CDT		$_{h_{\mathrm{c}}}$NR-CDT		$_{h_{\mathrm{d}}}$NR-CDT		$_{\mathrm{tv}}$NR-CDT	
	$\|\cdot\|_\infty$	$\|\cdot\|_2$	$\|\cdot\|_\infty$	$\|\cdot\|_2$	$\|\cdot\|_\infty$	$\|\cdot\|_2$	$\|\cdot\|_\infty$	$\|\cdot\|_2$	$\|\cdot\|_\infty$	$\|\cdot\|_2$	$\|\cdot\|_\infty$	$\|\cdot\|_2$	$\|\cdot\|_\infty$	$\|\cdot\|_2$	$\|\cdot\|_\infty$	$\|\cdot\|_2$
2	0.333	0.366	0.466	0.433	0.400	0.466	0.333	0.366	0.400	**0.500**	0.200	0.333	0.133	0.333	0.133	0.333
4			0.500	0.333	0.733	0.733	0.600	0.733	0.833	0.633	**0.900**	0.866	0.766	0.833	0.700	0.800
8			0.400	0.366	0.666	**0.933**	0.666	**0.933**	0.733	0.733	**0.933**	0.800	**0.933**	0.866	0.666	0.733
16			0.500	0.366	**1.000**	**1.000**	**1.000**	**1.000**	0.966	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.933	0.933
32			0.433	0.366	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
64			0.433	0.366	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
128			0.400	0.366	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
256			0.400	0.366	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 2 NT classification accuracies for academic dataset with 10 samples per class and different numbers of angles.

angles	Eucl.	R-CDT	$_{\mathrm{m}}$NR-CDT	$_{h_{\mathrm{d}}}$NR-CDT	$_{\mathrm{tv}}$NR-CDT
2	0.144	0.122	0.089	**0.122**	**0.122**
4		0.133	0.278	0.289	**0.311**
8		0.133	**0.289**	0.233	0.244
16		0.144	**0.567**	0.389	0.389
32		0.144	0.478	0.422	**0.611**
64		0.144	0.700	0.711	**0.944**
128		0.144	0.678	0.678	**1.000**
256		0.144	0.678	0.667	**1.000**

Table 3 NT classification accuracies for polygon dataset with 10 samples per class and varying numbers of angles.

reaches perfect results. This indicates its potential in real-world applications and may be attributed to the fact that $_{\mathrm{tv}}$NR-CDT uses the whole information over the entire range of angles in $[0, 2\pi)$, while $_{\mathrm{m}}$NR-CDT and $_{h_{\mathrm{d}}}$NR-CDT only consider the largest and smallest angle information.

5.1.2 Nearest Neighbour Classification

For a more realistic scenario, we consider the LinMNIST dataset with 100 samples per class. As in this case no templates exist, we replace NT classification by the nearest neighbour (1-NN) method, which assigns the label of the nearest member of a given training set with respect to the Euclidean distance in feature space. To this end, a subset of varying size is randomly selected to serve as training data and the remaining samples of the dataset are used for testing. This process is repeated 20 times and the mean and standard deviation (std) of the resulting classification accuracies are reported in Table 4, where we use (up to) 256 angles, 500 radii and 256 interpolation points.

While the Euclidean baseline and R-CDT perform at the level of random guessing, we observe that all our $_{h}$NR-CDT variants yield remarkable results in this limited data setting, that improve with increasing numbers of angles and selected training samples. In particular, when using only *one* training datum per class, $_{\mathrm{m}}$NR-CDT and

train	angle	Eucl.	R-CDT	$_{\mathrm{m}}$NR-CDT	$_{h_{\mathrm{d}}}$NR-CDT	$_{\mathrm{tv}}$NR-CDT
1	2	0.344±0.024	0.331±0.016	**0.477±0.083**	0.430±0.084	0.429±0.099
	4		0.336±0.028	**0.589±0.089**	**0.589±0.108**	0.571±0.135
	8		0.340±0.027	0.763±0.080	0.718±0.126	**0.769±0.130**
	16		0.340±0.027	0.796±0.122	0.768±0.114	**0.828±0.122**
	32		0.340±0.027	0.818±0.093	0.768±0.120	**0.830±0.119**
	64		0.340±0.027	0.820±0.090	0.760±0.127	**0.824±0.119**
	128		0.340±0.027	0.820±0.091	0.752±0.127	**0.823±0.119**
	256		0.340±0.027	0.820±0.091	0.752±0.127	**0.822±0.119**
5	2	0.340±0.038	0.335±0.022	0.541±0.047	**0.548±0.058**	0.530±0.047
	4		0.343±0.025	**0.717±0.051**	0.706±0.049	0.696±0.071
	8		0.343±0.032	0.832±0.037	**0.837±0.046**	0.829±0.041
	16		0.343±0.031	0.863±0.023	0.849±0.033	**0.879±0.017**
	32		0.344±0.031	0.885±0.019	0.865±0.026	**0.891±0.019**
	64		0.344±0.031	0.884±0.021	0.864±0.025	**0.889±0.021**
	128		0.344±0.031	0.886±0.019	0.864±0.024	**0.890±0.022**
	256		0.344±0.031	0.886±0.019	0.864±0.024	**0.890±0.023**
10	2	0.348±0.022	0.335±0.021	**0.586±0.021**	0.523±0.061	0.541±0.035
	4		0.350±0.028	**0.734±0.048**	0.713±0.038	0.729±0.027
	8		0.351±0.032	0.846±0.025	0.845±0.028	**0.856±0.017**
	16		0.352±0.032	0.870±0.019	0.866±0.026	**0.891±0.016**
	32		0.352±0.033	**0.899±0.021**	0.878±0.022	0.896±0.016
	64		0.352±0.033	0.899±0.020	0.878±0.018	**0.901±0.015**
	128		0.352±0.033	**0.902±0.020**	0.878±0.018	0.901±0.013
	256		0.352±0.033	**0.901±0.020**	0.880±0.018	**0.901±0.013**

Table 4 1-NN classification accuracies (mean±std) for LinMNIST dataset with 100 samples of classes $\{1, 5, 7\}$ and different numbers of training samples and angles.

$_{\mathrm{tv}}$NR-CDT already reach an accuracy of 82%, which improves to 90% for 10 training samples.

5.1.3 Support Vector Machines

In this last set of numerical experiments regarding image classification, we investigate the performance of our feature representations combined with linear support vector machines (SVMs) [43]. To this end, we utilize the academic dataset with 500 samples of classes $\{1, 2\}$ and the LinMNIST dataset with 500 samples of classes $\{1, 7\}$. We again randomly select a subset of varying size to train the SVMs and use the remaining samples of the dataset for testing.

The SVM classification accuracies (mean ± std based on 20 repetitions) for the academic dataset are reported in Table 5, where we use (up to) 256

train	angle	Eucl.	R-CDT	$_{\mathrm{m}}$NR-CDT	$_{h_{\mathrm{d}}}$NR-CDT	$_{\mathrm{tv}}$NR-CDT
1	2	0.500±0.015	0.499±0.033	0.617±0.126	0.663±0.119	**0.676±0.131**
	4		0.499±0.023	0.777±0.084	0.725±0.112	**0.783±0.102**
	8		0.520±0.018	**0.976±0.059**	0.913±0.094	0.763±0.139
	16		0.514±0.023	**1.000±0.000**	0.988±0.023	0.937±0.055
	32		0.578±0.020	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
	64		0.505±0.025	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
5	2	0.508±0.019	0.515±0.039	**0.863±0.052**	0.744±0.051	0.760±0.084
	4		0.522±0.029	**0.971±0.040**	0.947±0.041	0.964±0.031
	8		0.528±0.045	**1.000±0.000**	**1.000±0.000**	0.984±0.017
	16		0.556±0.058	**1.000±0.000**	**1.000±0.000**	0.986±0.011
	32		0.596±0.076	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
	64		0.599±0.086	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
10	2	0.518±0.023	0.549±0.031	**0.936±0.049**	0.797±0.047	0.847±0.049
	4		0.572±0.048	**0.989±0.022**	0.978±0.009	0.972±0.045
	8		0.630±0.050	**1.000±0.000**	**1.000±0.000**	0.991±0.013
	16		0.703±0.062	**1.000±0.000**	**1.000±0.000**	0.996±0.006
	32		0.767±0.092	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
	64		0.768±0.093	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
25	2	0.527±0.014	0.593±0.085	**0.975±0.025**	0.881±0.041	0.940±0.022
	4		0.730±0.068	**1.000±0.000**	0.986±0.007	0.990±0.007
	8		0.868±0.064	**1.000±0.000**	**1.000±0.000**	0.998±0.005
	16		0.975±0.037	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
	32		0.991±0.014	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
	64		0.985±0.022	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
50	2	0.531±0.017	0.632±0.031	0.983±0.010	0.933±0.022	**0.979±0.016**
	4		0.863±0.035	**1.000±0.000**	0.989±0.004	0.993±0.007
	8		0.971±0.023	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
	16		**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
	32		**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
	64		0.991±0.018	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**

Table 5 SVM classification accuracies (mean±std) for academic dataset with 500 samples of classes $\{1, 2\}$ and different numbers of training samples and Radon angles.

train	angle	Eucl.	R-CDT	$_{\mathrm{m}}$NR-CDT	$_{h_{\mathrm{d}}}$NR-CDT	$_{\mathrm{tv}}$NR-CDT
1	2	0.504±0.023	0.505±0.030	**0.678±0.102**	0.520±0.024	0.518±0.046
	4		0.496±0.027	**0.783±0.083**	0.584±0.109	0.657±0.099
	8		0.505±0.023	**0.877±0.120**	0.805±0.091	0.784±0.145
	16		0.502±0.025	**0.931±0.036**	0.827±0.181	0.837±0.099
	32		0.510±0.019	**0.957±0.010**	0.805±0.129	0.836±0.185
	64		0.519±0.025	**0.952±0.011**	0.872±0.120	0.903±0.058
5	2	0.504±0.018	0.515±0.028	**0.717±0.082**	0.528±0.047	0.535±0.031
	4		0.518±0.021	**0.821±0.064**	0.709±0.084	0.744±0.036
	8		0.517±0.042	**0.937±0.031**	0.873±0.042	0.869±0.042
	16		0.515±0.021	**0.949±0.021**	0.911±0.045	0.933±0.020
	32		0.514±0.044	**0.954±0.011**	0.944±0.018	0.931±0.033
	64		0.507±0.028	**0.961±0.007**	0.947±0.016	0.926±0.085
10	2	0.507±0.023	0.511±0.026	**0.784±0.032**	0.537±0.037	0.562±0.034
	4		0.514±0.026	**0.847±0.035**	0.753±0.022	0.745±0.045
	8		0.510±0.031	**0.944±0.013**	0.907±0.018	0.871±0.083
	16		0.528±0.032	**0.953±0.010**	0.935±0.016	0.937±0.029
	32		0.534±0.030	**0.960±0.010**	0.946±0.027	0.948±0.027
	64		0.541±0.041	**0.956±0.019**	0.955±0.020	0.949±0.033
25	2	0.517±0.025	0.521±0.024	**0.799±0.017**	0.569±0.034	0.581±0.027
	4		0.513±0.023	**0.860±0.028**	0.768±0.013	0.782±0.018
	8		0.533±0.027	**0.946±0.015**	0.909±0.010	0.923±0.030
	16		0.541±0.024	**0.959±0.011**	0.950±0.009	0.946±0.024
	32		0.550±0.037	**0.962±0.010**	0.960±0.012	0.955±0.022
	64		0.556±0.027	**0.962±0.013**	0.962±0.012	0.961±0.021
50	2	0.518±0.015	0.535±0.025	**0.812±0.021**	0.554±0.033	0.591±0.028
	4		0.528±0.022	**0.881±0.016**	0.780±0.008	0.793±0.011
	8		0.552±0.027	**0.957±0.008**	0.923±0.006	0.937±0.016
	16		0.567±0.032	**0.962±0.009**	0.855±0.005	0.956±0.016
	32		0.571±0.029	**0.968±0.007**	0.967±0.007	0.967±0.007
	64		0.583±0.032	**0.968±0.009**	0.967±0.009	**0.968±0.008**

Table 6 SVM classification accuracies (mean±std) for LinMNIST dataset with 500 samples of classes $\{1, 7\}$ and different numbers of training samples and Radon angles.

angles, 850 radii and 256 interpolation points. We observe that the Euclidean embedding performs on the level of random guessing regardless of the increasing number of training samples. R-CDT reaches perfect results, but needs a lot of training data. In contrast to this, all our $_{h}$NR-CDT representations yield perfect classification results already for *one* training sample per class, provided the number of Radon angles is sufficiently large.

The classification results for the LinMNIST dataset are reported in Table 6, where we use (up to) 256 angles, 500 radii and 256 interpolation points. Now, both the Euclidean baseline and R-CDT yield accuracies comparable to random guessing. As opposed to this, our new $_{h}$NR-CDT representations yield near perfect results, even in the very challenging case of only *one* training sample per class, with $_{\mathrm{m}}$NR-CDT performing best.

5.2 Image Clustering

While our experiments in the previous section are all based on a supervised learning setting, either by making use of known templates or choosing some part of the dataset as training data, we now transition to an unsupervised setting. We again consider the three datasets from before, now with 100 samples per class, and study unsupervised classification in the sense that we first perform a cluster analysis on a subset of the data and, thereon, classify the remaining part based on the corresponding cluster centres in feature space.

As our theory predicts linear separability in $_{h}$NR-CDT space, we apply the classical k-means algorithm to cluster the first 50 samples of each class in the respective feature space, where k is the expected number of classes. Thereon, we use the corresponding cluster centres to assign the remaining data to the closest cluster with respect to the Euclidean norm in feature space. The first step is referred to as training phase, the second as test phase. We evaluate the quality of the assignments by the so-called Rand index (RI) [44] and the variational information (VI) [45], which we now introduce in more details.

Let $[n] \coloneqq \{1, 2, ..., n\}$ and consider two partitions $\mathcal{U} = \{U_1, ..., U_k\}$ and $\mathcal{V} = \{V_1, ..., V_k\}$ of $[n]$. Then, the Rand index is defined as

$$\mathrm{RI}(\mathcal{U}, \mathcal{V}) \coloneqq \frac{a+b}{\binom{n}{2}}$$

with the number of accordance

$$a = |\{(x, y) \in [n]^2 \mid \exists\, s, t \in [k] \colon x, y \in U_s \cap V_t\}|$$

and the number of distinction

$$b = |\{(x, y) \in [n]^2 \mid \nexists\, s, t \in [k] \colon x, y \in U_s \cup V_t\}|.$$

Intuitively, RI measures the similarity between two clusterings and can be seen as prediction accuracy. To introduce the variational information, let

$$p_{st} \coloneqq \frac{|U_s \cap V_t|}{n}, \quad p_{s\cdot} \coloneqq \frac{|U_s|}{n}, \quad p_{\cdot t} \coloneqq \frac{|V_t|}{n}$$

for $(s, t) \in [k]^2$ and consider the entropies

$$H(\mathcal{U}) \coloneqq -\sum_{s=1}^{k} p_{s\cdot} \log(p_{s\cdot}),$$
$$H(\mathcal{V}) \coloneqq -\sum_{t=1}^{k} p_{\cdot t} \log(p_{\cdot t})$$

as well as the mutual information

$$I(\mathcal{U}, \mathcal{V}) \coloneqq \sum_{s,t=1}^{k} p_{st} \log\left(\tfrac{p_{st}}{p_{s\cdot} p_{\cdot t}}\right).$$

Then, the variational information is defined as

$$\mathrm{VI}(\mathcal{U}, \mathcal{V}) \coloneqq H(\mathcal{U}) + H(\mathcal{V}) - 2I(\mathcal{U}, \mathcal{V})$$

and quantifies the amount of information that is lost or gained when changing from $\mathcal{U}$ to $\mathcal{V}$. In our experiments, we compare the clustering induced by the true labels with the k-means clustering results. For visualization, we make use of 2d or 3d principal component analysis (PCA), where the true classes are visualized by different symbols (e.g. ◯, ☆, △, ...) and the k-means cluster centres by coloured boxes (e.g. ■, ■, ■, ...). The cluster members are coloured opaquely (e.g. ■, ■, ■, ...) and the test data transparently (e.g. ■, ■, ■, ...) according to the nearest cluster centre.

		Eucl.	R-CDT	${}_{\mathrm{m}}$NR-CDT	${}_{\mathrm{tv}}$NR-CDT
$\mathrm{RI}_{\mathrm{train}}$	(↑)	0.3378	0.5486	**1.0000**	**1.0000**
$\mathrm{VI}_{\mathrm{train}}$	(↓)	1.1492	2.1676	**0.0000**	**0.0000**
$\mathrm{RI}_{\mathrm{test}}$	(↑)	0.3288	0.5537	**1.0000**	**1.0000**
$\mathrm{VI}_{\mathrm{test}}$	(↓)	1.0986	2.1207	**0.0000**	**0.0000**

Table 7 Quality measures for 3-means clustering of the academic dataset with 100 images per class, where 50 images are used for training and the rest for testing.

		Eucl.	R-CDT	${}_{\mathrm{m}}$NR-CDT	${}_{\mathrm{tv}}$NR-CDT
$\mathrm{RI}_{\mathrm{train}}$	(↑)	0.1367	0.7406	0.8602	**0.9752**
$\mathrm{VI}_{\mathrm{train}}$	(↓)	2.2449	3.8193	1.6014	**0.2853**
$\mathrm{RI}_{\mathrm{test}}$	(↑)	0.1091	0.7130	0.8579	**0.9752**
$\mathrm{VI}_{\mathrm{test}}$	(↓)	2.1972	3.6659	1.6275	**0.2706**

Table 8 Quality measures for 9-means clustering of the polygon dataset with 100 images per class, where 50 images are used for training and the rest for testing.

In all our numerical experiments, we use 850 radii and 256 angles for the Radon transform and 256 interpolation points for the CDT. The results for the academic dataset are reported in Table 7. We observe that our ${}_{h}$NR-CDT feature representations clearly outperform the Euclidean baseline and R-CDT embedding, yielding perfect RI and VI values on both the train and test set. Moreover, the cluster visualization in Fig. 7 indicates that the classes are indeed mapped to single points in ${}_{h}$NR-CDT space, as predicted by our theory.

In case of the polygon dataset we observe that ${}_{\mathrm{tv}}$NR-CDT outperforms even the ${}_{\mathrm{m}}$NR-CDT with a near perfect Rand index of 97%, see Table 8. In contrast to this, ${}_{\mathrm{m}}$NR-CDT yields a Rand index of 86%, which is still significantly better than the Euclidean baseline and R-CDT. An inspection of the cluster visualization in Fig. 9 reveals that now the classes are not mapped to singletons in ${}_{h}$NR-CDT space. This, however, is to be expected as the polygon dataset includes small non-affine perturbations. Nevertheless, the classes are mapped to well-separated sets, which is more pronounced for ${}_{\mathrm{tv}}$NR-CDT than ${}_{\mathrm{m}}$NR-CDT. This again indicates the potential of including the whole angular information instead of only considering the maximum.

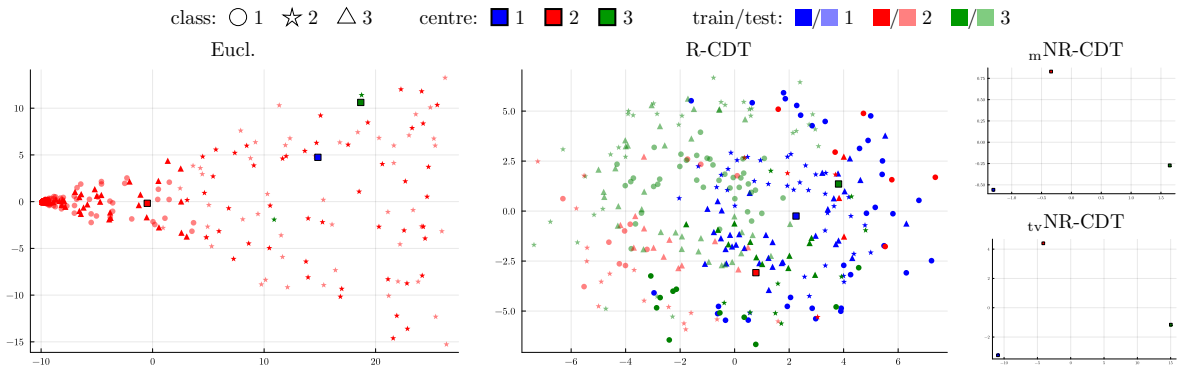

Fig. 7 3-means cluster visualization for the academic dataset using a 2d PCA in the respective feature space.

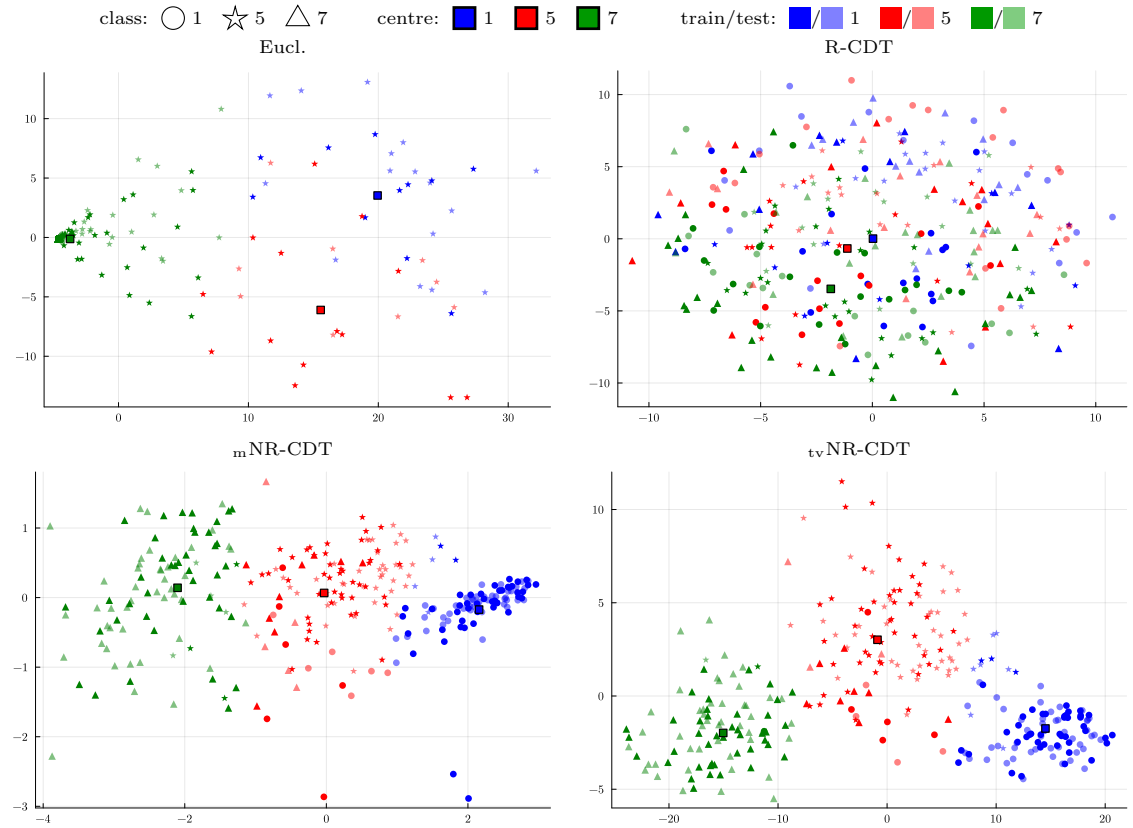

Fig. 8 3-means cluster visualization for the LinMNIST dataset using a 2d PCA in the respective feature space.

Finally, for the LinMNIST dataset $_{\mathrm{tv}}$NR-CDT and $_{\mathrm{m}}$NR-CDT yield comparable results and perform significantly better than the Euclidean and R-CDT embedding, which are nearly on the same level, cf. Table 9. The cluster visualizations in Fig. 8 show that $_{\mathrm{m}}$NR-CDT and $_{\mathrm{tv}}$NR-CDT yield nicely separated clusters with larger distances in $_{\mathrm{tv}}$NR-CDT space, which might explain the slightly better results for $_{\mathrm{tv}}$NR-CDT.

5.3 Multi-dimensional Classification

In the final set of numerical experiments, we leave the 2d Euclidean setting and explore the aptitude

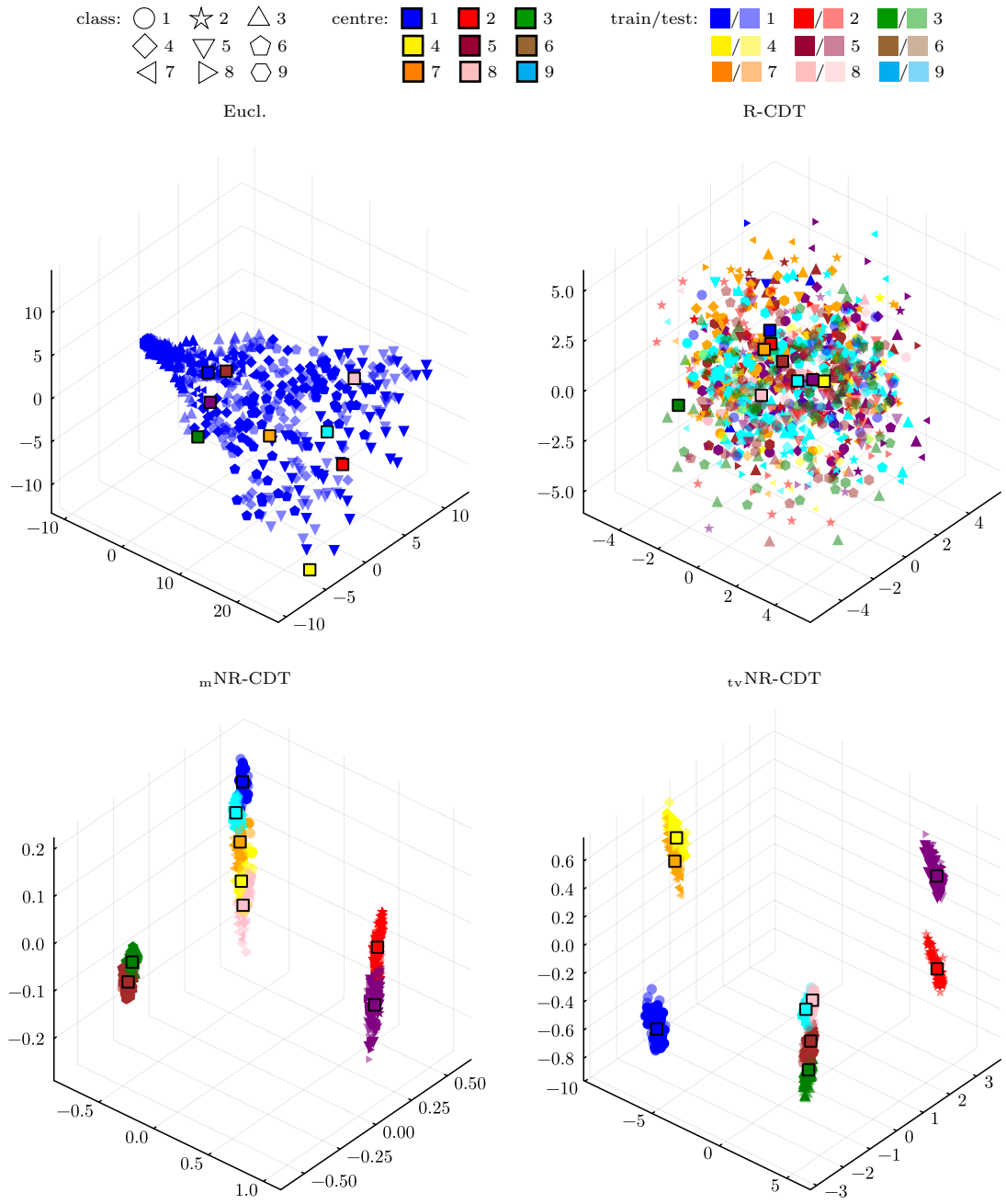

Fig. 9 9-means cluster visualization for the polygon dataset using a 3d PCA in the respective feature space.

of our generalized $_h$NR-CDTs regarding classification tasks on $\mathbb{X} = \mathbb{R}^3$ and $\mathbb{X} = \mathrm{SO}(3)$. For this, we consider classes of empirical measures

$$\mu := \frac{1}{K} \sum_{k=1}^{K} \delta_{\mathbf{x}_k}, \quad \mathbf{x}_k \in \mathbb{X},$$

whose generalized restricted Radon transforms are analytically given by

$$\mathcal{R}_{\phi,\boldsymbol{\theta}}[\mu] = \frac{1}{K} \sum_{k=1}^{K} \delta_{\phi(\mathbf{x}_k,\boldsymbol{\theta})}.$$

	Eucl.	R-CDT	$_{\mathrm{m}}$NR-CDT	$_{\mathrm{tv}}$NR-CDT
$\mathrm{RI}_{\mathrm{train}}$ (↑)	0.4832	0.5517	0.8385	**0.8526**
$\mathrm{VI}_{\mathrm{train}}$ (↓)	1.2223	2.1840	0.8303	**0.7978**
$\mathrm{RI}_{\mathrm{test}}$ (↑)	0.4826	0.5489	0.8488	**0.8798**
$\mathrm{VI}_{\mathrm{test}}$ (↓)	1.2255	2.1577	0.8293	**0.6713**

Table 9 Quality measures for 3-means clustering of the LinMNIST dataset with 100 images per class, where 50 images are used for training and the rest for testing.

Then, the corresponding quantile functions are piecewise constant and these are sampled on an equispaced grid to calculate the final $_h$NR-CDTs.

5.3.1 3D Object Recognition

The first generalized $_h$NR-CDT we study numerically is the multi-dimensional extension in § 4.1, which can be applied to the recognition of 3d objects. For first proof-of-concept experiments, we consider the following two datasets:

a) **Animal dataset**. We use the 3d triangular mesh information of six animals—lion, cat, camel, horse, flamingo, elephant—provided in [46, 47], where each mesh consists of 5000 vertices. More precisely, we equip these vertices with a uniform probability distribution to obtain empirical measures on $\mathbb{R}^3$. On the basis of these six templates, the dataset itself is generated by applying 10 random affine transformations per template, consisting of anisotropic scaling factors in $[0.5, 1.0]$ and shearing in $[-15°, 15°]$ for each coordinate direction as well as random 3d rotations and shifts in $[-25, 25]^3$. The animal dataset is illustrated in Figure 10.

b) **ABC dataset**. The original 'A Big CAD Model Dataset' [48] consists of over 1 million geometrical models. To build our affine dataset, we use a subset of 80 shapes that are visualized in Figure 11. Each object mesh consists of 2028 vertices, which we equip with the uniform probability distribution to obtain an empirical measure in $\mathbb{R}^3$. Similar to the animal dataset, each template is used to construct a class of 10 random affinely transformed versions. For the construction, we again employ anisotropic scaling factors in $[0.5, 1.0]$ and shearing in $[-15°, 15°]$ as well as random 3d rotations and shifts in $[-25, 25]^3$.

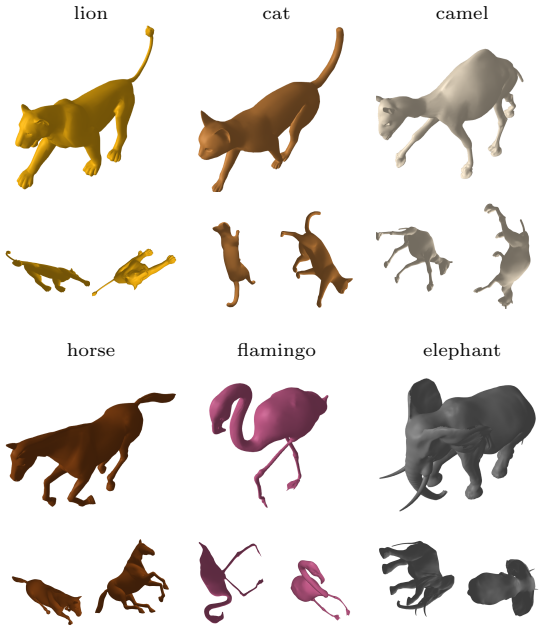

Fig. 10 Visualization of the animal dataset: template meshes (top) and affinely transformed samples (bottom).

Fig. 11 Visualization of the templates of ABC dataset.

Note that the main crucial point in the multi-dimensional setting compared to the 2d setting is that h in the definition of the generalized $_h$NR-CDT operates on a larger set of directions, here $\Theta = \mathbb{S}_2$. Figuratively, this means that more information is condensed, potentially lowering the feature extraction capabilities of the $_h$NR-CDT. Considering the $_h$NR-CDT of the six animal classes in Figure 12, where the underlying CDTs are sampled on an equispaced grid of 1000 points in $(0, 1)$, we notice that this apprehension does

directions	R-CDT	$_{\mathrm{m}}$NR-CDT	$_{h_{\mathrm{a}}}$NR-CDT	$_{h_{\mathrm{d}}}$NR-CDT
2	0.233	0.450	**0.450**	0.267
4	0.233	0.283	**0.417**	0.367
8	0.233	0.733	**0.767**	0.650
16	0.250	0.883	0.950	**0.967**
32	0.250	0.933	**1.000**	0.967
64	0.250	0.983	**1.000**	**1.000**
128	0.250	**1.000**	**1.000**	**1.000**
256	0.250	**1.000**	**1.000**	**1.000**

Table 10 NT classification accuracies for the animal dataset with 10 samples per class and varying numbers of directions (in form of Fibonacci-like points on $\mathbb{S}_2$).

directions	R-CDT	$_{\mathrm{m}}$NR-CDT	$_{h_{\mathrm{a}}}$NR-CDT	$_{h_{\mathrm{d}}}$NR-CDT
2	0.030	0.228	**0.239**	0.089
4	0.028	0.346	**0.351**	0.218
8	0.023	0.505	**0.538**	0.389
16	0.025	0.710	**0.745**	0.604
32	0.026	0.746	**0.754**	0.676
64	0.026	0.829	**0.829**	0.775
128	0.026	0.850	**0.859**	0.785
256	0.026	**0.975**	0.894	0.846
512	0.026	0.913	**0.925**	0.888
1024	0.026	**0.919**	0.915	0.909
2048	0.026	**0.945**	0.936	0.928

Table 11 NT classification accuracies for the ABC dataset with 10 samples per class and varying numbers of directions (in form of Fibonacci-like points on $\mathbb{S}_2$).

not materialize. Again, as suggested by the theory, each animal class shrinks to a single point in $_h$NR-CDT space, and the different $_h$NR-CDTs are well distinguishable. For the numerical computations, we rely on the Fibonacci-like points to discretize the direction set $\mathbb{S}_2$ in style of [49], given by

$$\boldsymbol{\theta}_k := \begin{bmatrix} \sqrt{1-z_k}\cos(k\varphi) \\ \sqrt{1-z_k}\sin(k\varphi) \\ z_k \end{bmatrix}, \quad k \in \{0, ..., n-1\},$$

with $z_k = 1 - \frac{2k+1}{n}$ and golden angle $\varphi = \pi(3 - \sqrt{5})$. Note that we do not consider the $_{\mathrm{tv}}$NR-CDT in this setting, since there is no straightforward generalization to functions on the sphere $\mathbb{S}_2$.

We adapt the NT classification experiments from § 5.1.1, i.e., we label the affinely transformed data based on the nearest template in $_h$NR-CDT space with respect to the Euclidean norm. The achieved classification accuracies for the animal dataset are reported in Table 10. As expected, all employed $_h$NR-CDT variants yield perfect classifications, whereas the multi-dimensional R-CDT cannot classify the different animals. Worth mentioning, we obtain already convincing results for small numbers of Fibonacci points.

The NT classification accuracies for the ABC dataset are shown in Table 11. Here, the main challenge is the large number of 80 different classes, which are partly based on very similar templates. Nevertheless, for a sufficient fine discretization of $\mathbb{S}_2$, most samples are correctly classified when using our $_h$NR-CDT feature representations. Note again that R-CDT is not designed to distinguish affine classes based on anisotropic scaling and shearing used in this experiment.

In conclusion, these first proof-of-concept simulations indicate the usefulness of our multi-dimensional $_h$NR-CDT feature representations in 3d point cloud recognition tasks. Further experiments on corresponding real-world applications and sensitivity studies are left for future research.

5.3.2 Classification on the 3D Rotation Group

In our last set of numerical experiments we leave the Euclidean setting and study a classification task on the 3d rotation group SO(3) as in § 4.3. To this end, we consider a synthetic dataset based on three empirical template measures, each supported on 1000 rotation matrices. More precisely, the matrices for the first template are generated by the Matrix-Fisher distribution [50] around I_3 with concentration $\kappa = 10$, the matrices for the second template are uniformly distributed on the subset of rotations around the x-y-equator, and the matrices for the third template are constructed by computing the QR decomposition of random Gaussian matrices in $\mathbb{R}^{3\times3}$. Thereon, these template measures are transformed by applying random rotations to form the dataset, we refer to as **rotation dataset**, see Fig. 14 for visualization.

For the numerical computations, we need to discretize the direction set, which is now given by $\Theta = \mathrm{SO}(3)$. For this, we make use of the so-called Super-Fibonacci points [51], defined by the Euler parameters

$$a_k = \sqrt{\tfrac{2k+1}{2n}}\sin\big(\tfrac{\pi(2k+1)}{\varphi_1}\big),\ c_k = \sqrt{1-\tfrac{2k+1}{2n}}\cos\big(\tfrac{\pi(2k+1)}{\varphi_2}\big),$$
$$b_k = \sqrt{\tfrac{2k+1}{2n}}\cos\big(\tfrac{\pi(2k+1)}{\varphi_1}\big),\ d_k = \sqrt{1-\tfrac{2k+1}{2n}}\sin\big(\tfrac{\pi(2k+1)}{\varphi_2}\big),$$

for $k \in \{0, \dots, n-1\}$, where $\varphi_1 = \sqrt{2}$ and φ_2 is the positive real solution to $\varphi_2^4 = \varphi_2 + 4$. One can show that $\varphi_2 = \frac{1}{2}(\sqrt{m} + \sqrt{2/\sqrt{m} - m})$ with

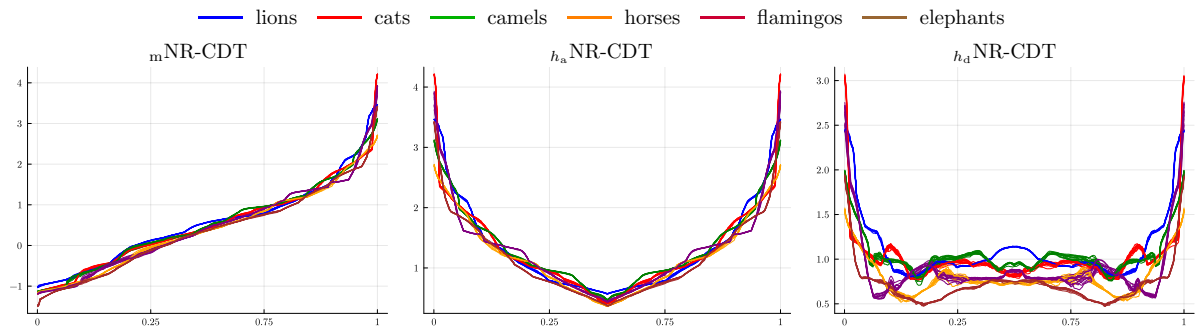

Fig. 12 Visualization of various $_h$NR-CDTs of the animal dataset with 10 samples per class and 2048 Fibonacci-like points.

rotations	R-CDT	$_{\mathrm{m}}$NR-CDT	$_{h_{\mathrm{a}}}$NR-CDT	$_{h_{\mathrm{d}}}$NR-CDT
2	0.633	0.933	0.933	0.433
4	0.400	**1.000**	**1.000**	0.733
8	0.333	**1.000**	**1.000**	0.900
16	0.367	**1.000**	**1.000**	0.967
32	0.367	**1.000**	**1.000**	**1.000**
64	0.400	**1.000**	**1.000**	**1.000**

Table 12 NT classification accuracies for rotation dataset with 10 samples per class and varying numbers of rotations (Super-Fibonacci points on SO(3)).

$m = \sqrt[3]{m_+} + \sqrt[3]{m_-}$ and $m_\pm = \frac{1}{18}(9 \pm \sqrt{49233})$. The rotation matrices $\boldsymbol{\theta}_k \in \mathrm{SO}(3)$ are then given by the Euler–Rodrigues formula

$$\boldsymbol{\theta}_k := \begin{bmatrix} 1-2(c_k^2+d_k^2) & 2(b_kc_k-a_kd_k) & 2(b_kd_k+a_kc_k) \\ 2(b_kc_k+a_kd_k) & 1-2(b_k^2+d_k^2) & 2(c_kd_k-a_kb_k) \\ 2(b_kd_k-a_kc_k) & 2(c_kd_k+a_kb_k) & 1-2(b_k^2+c_k^2) \end{bmatrix}.$$

We again adapt the NT classification experiments from § 5.1.1 and label the transformed data based on the nearest template in $_h$NR-CDT space with respect to the Euclidean norm. The $_h$NR-CDTs of the entire dataset are depicted in Fig. 13, where the underlying CDTs are sampled on an equispaced grid consisting of 1000 points in $(0, 1)$, and show an almost perfect alignment within each class. The achieved classification accuracies are reported in Table 12. As expected, all employed $_h$NR-CDT variants yield perfect results, whereas R-CDT cannot distinguish the different classes.

6 Conclusion

In this paper, we introduced the novel generalized $_h$NR-CDT for feature representation and proved linear separability of classes generated by affine transforms of given templates. This was validated by numerical experiments in two- and three-dimensional settings showing a significant increase in classification accuracy over the original R-CDT. Potential future directions include the investigation of $_h$NR-CDT performance in combination with more advanced classification and clustering methods, like in [28, 52, 53], as well as a more in-depth numerical study in various real-world applications.

References

[1] Kantorovich, L.V.: On the translocation of masses. Journal of Mathematical Sciences **133**(4), 1381–1382 (2006) https://doi.org/10.1007/s10958-006-0049-2

[2] Villani, C.: Topics in Optimal Transportation. American Mathematical Society, Providence (2003). https://doi.org/10.1090/gsm/058

[3] Santambrogio, F.: Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20828-2

[4] Bogachev, V.I., Kolesnikov, A.V.: The Monge–Kantorovich problem: achievements, connections, and perspectives. Russian Mathematical Surveys **67**(5), 785–890 (2012) https://doi.org/10.1070/RM2012v067n05ABEH004808

[5] Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 1–9 (2013)

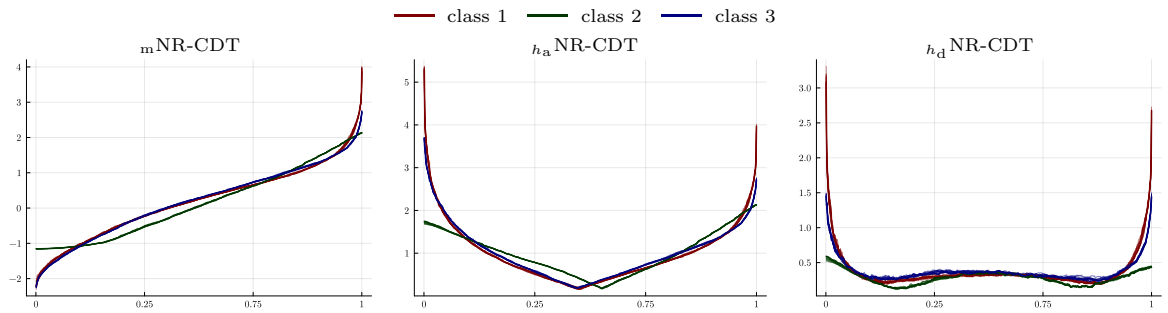

Fig. 13 Visualization of $_h$NR-CDTs of the rotation dataset with 10 samples per class and 2048 Super-Fibonacci points.

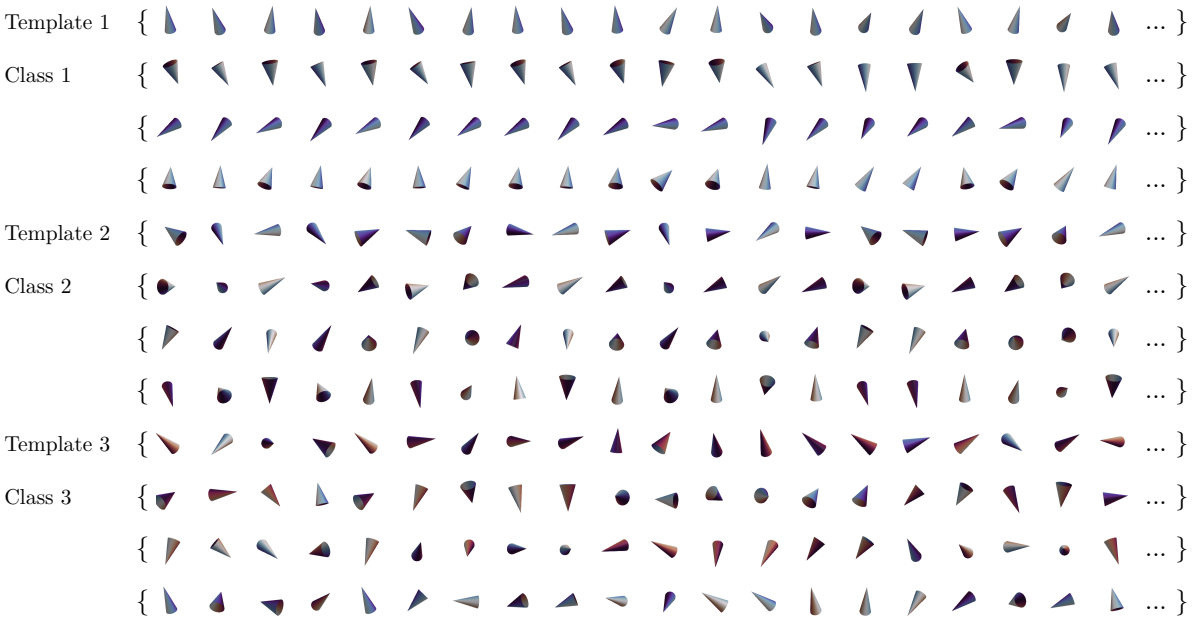

Fig. 14 Illustration of the rotation dataset by visualizing the action of the underlying rotation matrices on a cone pointing towards the north pole, whose surface is coloured by a periodic colour bar to depict rotations around its z-axis.

[6] Wang, W., Slepčev, D., Basu, S., Ozolek, J.A., Rohde, G.K.: A linear optimal transportation framework for quantifying and visualizing variations in sets of images. International Journal of Computer Vision **101**(2), 254–269 (2013) https://doi.org/10.1007/s11263-012-0566-z

[7] Park, S.R., Kolouri, S., Kundu, S., Rohde, G.K.: The cumulative distribution transform and linear pattern classification. Applied and Computational Harmonic Analysis **45**(3), 616–641 (2018) https://doi.org/10.1016/j.acha.2017.02.002

[8] Moosmüller, C., Cloninger, A.: Linear optimal transport embedding: provable Wasserstein classification for certain rigid transformations and perturbations. Information and Inference: A Journal of the IMA **12**(1), 363–389 (2023) https://doi.org/10.1093/imaiai/iaac023

[9] Bonneel, N., Rabin, J., Peyré, Pfister, H.: Sliced and Radon Wasserstein barycenters of measures. Journal of Mathematical Imaging and Vision **51**(1), 22–45 (2015) https://doi.org/10.1007/s10851-014-0506-3

[10] Shifat-E-Rabbi, M., Zhuang, Y., Li, S., Rubaiyat, A.H.M., Yin, X., Rohde, G.K.:

Invariance encoding in sliced-Wasserstein space for image classification with limited training data. Pattern Recognition **137**, 109268 (2023) https://doi.org/10.1016/j.patcog.2022.109268

[11] Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., Rohde, G.: Generalized sliced Wasserstein distances. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 261–272 (2019)

[12] Piening, M., Beinert, R.: Slicing the Gaussian mixture Wasserstein distance. Transactions on Machine Learning Research, 1–24 (2025)

[13] Grave, E., Joulin, A., Berthet, Q.: Unsupervised alignment of embeddings with Wasserstein Procrustes. In: International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1880–1890 (2019)

[14] Adamo, D., Corneli, M., Vuillien, M., Vila, E.: An in depth look at the Procrustes–Wasserstein distance: properties and barycenters. arXiv:2507.00894 (2025). https://doi.org/10.48550/arXiv.2507.00894

[15] Mémoli, F.: Gromov-Wasserstein distances and the metric approach to object matching. Foundations of Computational Mathematics **11**(4), 417–487 (2011) https://doi.org/10.1007/s10208-011-9093-5

[16] Beier, F., Piening, M., Beinert, R., Steidl, G.: Joint metric space embedding by unbalanced optimal transport with Gromov-Wasserstein marginal penalization. In: International Conference on Machine Learning (ICML) (2025)

[17] Salmona, A., Desolneux, A., Delon, J.: Gromov-Wasserstein-like distances in the Gaussian mixture models space. Transactions on Machine Learning Research (2024)

[18] Sejourne, T., Vialard, F.-X., Peyré, G.: The unbalanced Gromov–Wasserstein distance: conic formulation and relaxation. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 8766–8779 (2021)

[19] Beier, F., Beinert, R., Steidl, G.: On a linear Gromov–Wasserstein distance. IEEE Transactions on Image Processing **31**, 7292–7305 (2022) https://doi.org/10.1109/TIP.2022.3221286

[20] Beinert, R., Heiss, C., Steidl, G.: On assignment problems related to Gromov–Wasserstein distances on the real line. SIAM Journal on Imaging Sciences **16**(2), 1028–1032 (2023) https://doi.org/10.1137/22M1497808

[21] Vayer, T., Flamary, R., Courty, N., Tavenard, R., Chapel, L.: Sliced Gromov-Wasserstein. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 1–11 (2019)

[22] Piening, M., Beinert, R.: A novel sliced fused Gromov-Wasserstein distance. arXiv:2508.02364 (2025). https://doi.org/10.48550/arXiv.2508.02364

[23] Kolouri, S., Park, S.R., Rohde, G.K.: The Radon cumulative distribution transform and its application to image classification. IEEE Transactions on Image Processing **25**(2), 920–934 (2016) https://doi.org/10.1109/TIP.2015.2509419

[24] Ramm, A.G., Katsevich, A.I.: The Radon Transform and Local Tomography. CRC Press, Boca Raton (1996). https://doi.org/10.1201/9781003069331

[25] Natterer, F.: The Mathematics of Computerized Tomography. SIAM, Philadelphia (2001). https://doi.org/10.1137/1.9780898719284

[26] Kolouri, S., Park, S.R., Thorpe, M., Slepcev, D., Rohde, G.K.: Optimal mass transport. IEEE Signal Processing Magazine **34**(4), 43–59 (2017) https://doi.org/10.1109/MSP.2017.2695801

[27] Diaz Martin, R., Medri, I.V., Rohde, G.K.: Data representation with optimal transport. arXiv:2406.15503 (2024). https://doi.org/10.48550/arXiv.2406.15503

[28] Shifat-E-Rabbi, M., Yin, X., Rubaiyat, A.H.M., Li, S., Kolouri, S., Aldroubi,

A., Nichols, J.M., Rohde, G.K.: Radon cumulative distribution transform subspace modeling for image classification. Journal of Mathematical Imaging and Vision **63**, 1185–1203 (2021) https://doi.org/10.1007/s10851-021-01052-0

[29] Quellmalz, M., Beinert, R., Steidl, G.: Sliced optimal transport on the sphere. Inverse Problems **39**(10), 105005 (2023) https://doi.org/10.1088/1361-6420/acf156

[30] Quellmalz, M., Buecher, L., Steidl, G.: Parallelly sliced optimal transport on spheres and on the rotation group. Journal of Mathematical Imaging and Vision **66**(6), 951–976 (2024) https://doi.org/10.1007/s10851-024-01206-w

[31] Hauser, D., Beckmann, M., Koliander, G., Stiehl, H.S.: On image processing and pattern recognition for thermograms of watermarks in manuscripts – a first proof-of-concept. In: International Conference on Document Analysis and Recognition (ICDAR), pp. 91–107 (2024). https://doi.org/10.1007/978-3-031-70543-4_6

[32] Beckmann, M., Beinert, R., Bresch, J.: Max-normalized Radon cumulative distribution transform for limited data classification. In: International Conference on Scale Space and Variational Methods in Computer Vision (SSVM), pp. 241–254 (2024). https://doi.org/10.1007/978-3-031-92366-1_19

[33] Beckmann, M., Beinert, R., Bresch, J.: Normalized Radon Cumulative Distribution Transforms for Invariance and Robustness in Optimal Transport Based Image Classification. arXiv:2506.08761 (2025). https://doi.org/10.48550/arXiv.2506.08761

[34] Gel'fand, I.M., Graev, M.I., Shapiro, Z.Ya.: Differential forms and integral geometry. Functional Analysis and Its Applications **3**(2), 101–114 (1969) https://doi.org/10.1007/BF01674015

[35] Beylkin, G.: The inversion problem and applications of the generalized Radon transform. Communications on Pure and Applied Mathematics **37**(5), 579–599 (1984) https://doi.org/10.1002/cpa.3160370503

[36] Quinto, E.T.: The dependence of the generalized Radon transform on defining measures. Transactions of the American Mathematical Society **257**(2), 331–346 (1980) https://doi.org/10.1090/S0002-9947-1980-0552261-8

[37] Homan, A., Zhou, H.: Injectivity and stability for a generic class of generalized Radon transforms. Journal of Geometric Analysis **27**(2), 1515–1529 (2017) https://doi.org/10.1007/s12220-016-9729-4

[38] Kuchment, P.: Generalized transforms of Radon type and their applications. In: The Radon Transform, Inverse Problems, and Tomography. Proceedings of Symposia in Applied Mathematics, vol. 63, pp. 67–91. American Mathematical Society, Providence (2006). https://doi.org/10.1090/psapm/063/2208237

[39] Casadio Tarabusi, E., Picardello, M.A.: Radon transforms in hyperbolic spaces and their discrete counterparts. Complex Analysis and Operator Theory **15**(1), 13 (2021) https://doi.org/10.1007/s11785-020-01055-6

[40] Rustamov, R.M., Majumdar, S.: Intrinsic sliced Wasserstein distances for comparing collections of probability distributions on manifolds and graphs. In: International Conference on Machine Learning (ICML), pp. 29388–29415 (2023)

[41] Beckmann, M., Heilenkötter, N.: Equivariant neural networks for indirect measurements. SIAM Journal on Mathematics of Data Science **6**(3), 579–601 (2024) https://doi.org/10.1137/23M1582862

[42] Deng, L.: The MNIST database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine **29**(6), 141–142 (2012) https://doi.org/10.1109/MSP.2012.2211477

[43] Fan, R.-E., Chang, K.-W., Hsieh, C.-J.,

Wang, X.-R., Lin, C.-J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9**, 1871–1874 (2008). Software available at https://www.csie.ntu.edu.tw/~cjlin/liblinear/

[44] Hubert, L.J., Arabie, P.: Comparing partitions. Journal of Classification **2**, 193–218 (1985) https://doi.org/10.1007/BF01908075

[45] Meilă, M.: Comparing clusterings by the variation of information. In: Conference on Computational Learning Theory and Kernel Workshop (COLT/Kernel), pp. 173–187 (2003). https://doi.org/10.1007/978-3-540-45167-9_14

[46] Sumner, R.W., Popovic, J.: Mesh Data from Deformation Transfer for Triangle Meshes. (Access: 27th August 2025). http://people.csail.mit.edu/sumner/research/deftransfer/data.html

[47] Sumner, R.W., Popović, J.: Deformation transfer for triangle meshes. ACM Transactions on Graphics **23**(3), 399–405 (2004) https://doi.org/10.1145/1015706.1015736

[48] Koch, S., Matveev, A., Jiang, Z., Williams, F., Artemov, A., Burnaev, E., Alexa, M., Zorin, D., Panozzo, D.: ABC: A big CAD model dataset for geometric deep learning. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9593–9603 (2019). https://doi.org/10.1109/CVPR.2019.00983

[49] González, A.: Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices. Mathematical Geosciences **42**, 49–64 (2010) https://doi.org/10.1007/s11004-009-9257-x

[50] Downs, T.D.: Orientation statistics. Biometrika **59**(3), 665–676 (1972) https://doi.org/10.1093/biomet/59.3.665

[51] Alexa, M.: Super-Fibonacci spirals: Fast, low-discrepancy sampling of SO(3). In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8281–8290 (2022). https://doi.org/10.1109/CVPR52688.2022.00811

[52] Auffenberg, J., Bresch, J., Melnyk, O., Steidl, G.: Unsupervised Ground Metric Learning. arXiv:2507.13094 (2025). https://doi.org/10.48550/arXiv.2507.13094

[53] Xing, E., Jordan, M., Russell, S.J., Ng, A.: Distance metric learning with application to clustering with side-information. In: Advances in Neural Information Processing Systems (NeurIPS) (2002)