Bispectral OT: Dataset Comparison using Symmetry-Aware Optimal Transport

Editors: List of editors' names

Keywords: Optimal Transport, Discrete Optimal Transport, Dataset Comparison, Geometric Dataset Distance, Bispectrum, Symmetry Invariance, G-Bispectrum

1. Introduction

Optimal Transport is a widely used technique for distribution alignment, providing a rigorous framework to learn a transport plan that moves mass from one distribution to match another. It is playing an increasing role in machine learning, due to applications in transfer learning (Alvarez-Melis and Fusi, 2020), domain adaptation (Courty et al., 2017), generative modeling (Arjovsky et al., 2017; Bousquet et al., 2017), generalization bounds (Chuang et al., 2021), and imitation learning (Luo et al., 2023). In symmetry-rich settings, however, OT alignments based solely on pairwise geometric distances can ignore the intrinsic coherence structure of the data (e.g., labels in supervised settings). For example, in an image dataset with rotational symmetries, naive OT on the raw features can match images based upon orientation, rather than the shape of the object depicted. Thus, we seek a transport plan that is *symmetry-aware*: one that compares distributions in a way that is invariant to natural transformations of the data while retaining selectivity to informative structure.

We introduce *Bispectral Optimal Transport (BOT)*, a symmetry-aware extension of discrete OT that compares elements through their representation under the *bispectrum*—a complete invariant from group Fourier analysis that simultaneously encodes signal structure and invariance to group actions (Kakarala, 1993), first introduced to machine learning by Kondor (2007). By computing couplings in this bispectral embedding, BOT produces correspondences that are invariant to transformations induced by symmetry groups acting on the data, without discarding discriminative information. To the authors knowledge, this is the first work that encodes symmetry awareness into OT, with a more extended discussion of related work included in Appendix A. We demonstrate that BOT better preserves label information than standard OT on datasets augmented via symmetry transformations, effectively encoding relevant symmetries in the learned transport plans. This is critical in settings where the symmetries acting on the two distributions do not align: for example, matching images captured by cameras at different poses or orientations. In such cases, we want the comparison to reflect the distribution of semantic labels rather than the distribution of camera angles; BOT achieves this and substantially outperforms standard OT on datasets perturbed by synthetic transformations.

2. Background

Optimal Transport Optimal transport provides an elegant mathematical framework for aligning probability distributions (Villani et al., 2008). At a high level, OT seeks to transfer probability mass between distributions while minimizing a cost function of transportation,

yielding a notion of distance from the coupling. We are interested in the problem's discrete formulation, which considers two finite collections of points $\{\mathbf{x}^{(i)}\}_{i=1}^n \in \mathcal{X}^n$ and $\{\mathbf{y}^{(j)}\}_{j=1}^m \in \mathcal{Y}^n$ represented as empirical distributions: $\mu = \sum_{i=1}^n \mathbf{p}_i \delta_{\mathbf{x}^{(i)}}, \nu = \sum_{j=1}^m \mathbf{q}_j \delta_{\mathbf{y}^{(j)}}$ where $\mathbf{p}$ and $\mathbf{q}$ are probability vectors (non-negative and sum to one). Given a transportation cost $C \in \mathbb{R}^{n\times m}$ (also known as the *ground metric*) between pairs of points (e.g., $C_{ij} = \|x_i - y_j\|$), OT finds a correspondence between μ and ν that minimizes this cost. Formally, Kantorovich's formulation of optimal transport (Kantorovich, 1942) finds a coupling $\Gamma \in \mathbb{R}^{n\times m}$ that solves

$$\mathrm{OT}_c(\mu,\nu) = \min_{\Gamma\in\mathbb{R}_+^{n\times m}} \langle \Gamma, C\rangle \quad \text{s.t.} \quad \Gamma\mathbf{1} = \mathbf{p},\ \Gamma^\top\mathbf{1} = \mathbf{q}. \tag{1}$$

This coupling Γ can be interpreted as a soft matching between elements of μ and ν, in the sense that Γ_{ij} is high if $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(j)}$ are in correspondence, and low otherwise. In practice, the entropic-regularized Sinkhorn distance (Cuturi, 2013) is widely used, as it can be solved more efficiently via the Sinkhorn-Knopp algorithm. Specifically, the Sinkhorn distance solves $\min_{\Gamma\in\Pi(\mathbf{p},\mathbf{q})}\langle\Gamma, C\rangle - \epsilon H(\Gamma)$, where $H(\Gamma) = -\sum_{ij}\Gamma_{ij}\log\Gamma_{ij}$ is an entropic regularizer that smooths the transportation plan.

Bispectrum and Group-Invariant Fourier Embeddings We encode symmetries via the bispectrum, a Fourier invariant that is *complete*, removing specified group actions while preserving relative phase structure. To do this, we describe the group Fourier Transform (Rudin, 2017), with additional background deferred to Appendix B. Let $f : G \to \mathbb{C}$ be a signal on a group G with set of irreducible representations $Irr(G)$. The *Generalized Fourier Transform (GFT)* is the linear map $f \mapsto \hat{f}$, where the Fourier frequencies are indexed by $\rho \in Irr(G)$, defined in the discrete case as

$$\hat{f}_\rho = \sum_{g\in G} f(g)\rho(g). \tag{2}$$

In particular, for $G = \mathbb{Z}/n\mathbb{Z}$, the $\rho(g)$ are the n discrete Fourier frequencies: $\rho_k(g) = e^{-\mathbf{i}2\pi kg/n}$, for $k = 0,\dots,n-1$ and $g \in \mathbb{Z}/n\mathbb{Z}$. For translation by $t \in G$, defined as $f^t(x) := f(t^{-1}x)$, the GFT, like the classical FT, obeys the *Fourier shift property*: $\hat{f}^t_\rho = \rho(t)\hat{f}_\rho$. This equivariance is exploited in the well-known *power spectrum* $q_\rho = \hat{f}_\rho^\dagger \hat{f}_\rho$, which is invariant to translation but discards relative phase, losing structural information.

The *bispectrum* is a lesser known Fourier invariant restoring this structure. For 1D signals $f \in \mathbb{R}^n$ with Fourier coefficients $\hat{f} = (\hat{f}_0,\dots,\hat{f}_{n-1})$, the translation-invariant *bispectrum* is the complex matrix $B \in \mathbb{C}^{n\times n}$ with entries

$$B_{i,j} = \hat{f}_i\hat{f}_j\hat{f}^\dagger_{i+j \pmod n}, \tag{3}$$

which is invariant to phase shifts due to translation, but does so while preserving the signal's relative phase structure. For compact commutative groups, the bispectrum is defined as

$$B_{\rho_i,\rho_j} = \hat{f}_{\rho_i}\hat{f}_{\rho_j}\hat{f}_{\rho_i\rho_j}. \tag{4}$$

The non-commmutative bispectrum (Appendix C) is defined analogously, but accounts for matrix-valued irreps. Foundational work (Kakarala, 1993) shows that the bispectrum is the lowest-degree spectral invariant that is *complete*: it factors out specified group actions without losing signal structure. In labeled image datasets, this yields invariance to natural transformations (e.g., rotations) while preserving shape information—exactly the behavior required for symmetry-aware OT.

3. Bispectral OT

The Bispectral OT framework combines the symmetry-invariant properties of the bispectrum with the alignment capabilities of OT. The key idea is to compute an OT plan on symmetry-aware bispectral features, ensuring that the resulting correspondence respects the symmetries of the underlying data. Concretely, we propose a framework to construct the bispectral representation of grids (e.g., images) acted on by $SO(2)$, the group of planar rotations. For many non-commutative groups like $SO(3)$, bispectral feature embeddings as described in Kondor (2007) become computationally prohibitive for OT settings, which require us to compute and store a pairwise cost matrix between all bispectral features. The selective G-bispectrum (Mataigne et al., 2024) reduces this complexity, but it is unclear whether distances between its compressed features are expressive enough of global geometry for OT. Balancing efficiency with faithful geometry remains the central challenge. The rotation-invariant features are constructed as follows:

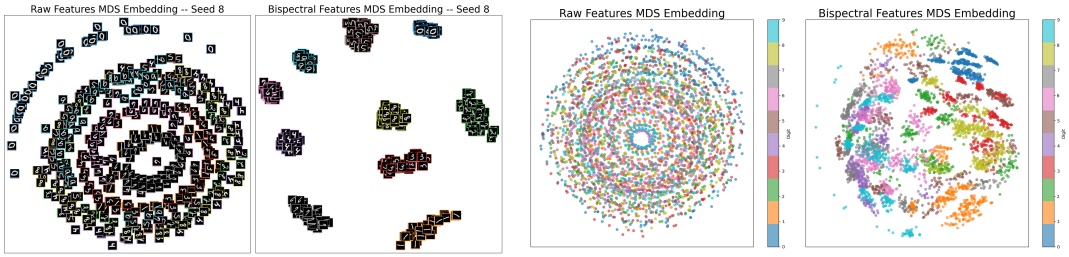
1. Convert each $M \times N$ pixel image into a discretized polar representation of size $R \times K$ (where R is the number of radial bins and K the number of angular bins).
2. For each fixed radius r, extract the $1 \times K$ angular slice and compute its 1D discrete Fourier transform (DFT) along the angular dimension $(f_r(\theta_1), \dots, f_r(\theta_K))$.
3. Compute the bispectrum of each slice. Since cyclic shifts of the angular dimension correspond to actions by $\mathbb{Z}/K\mathbb{Z}$ (i.e. discrete rotations), the bispectrum provides invariance to such transformations. Concretely, this maps our $R \times K$ polar image to $\mathbb{C}^{R \times K \times K}$ by mapping each DFT
$$(f_r(\theta_1), \dots, f_r(\theta_K)) \mapsto (f_r(\theta_i) f_r(\theta_j) f_r(\theta_i + \theta_j)^\dagger)_{i,j \in 1,\dots,K}.$$
4. Concatenate the bispectral features across radii to obtain a global $SO(2)$-invariant representation of the image.

These bispectral representations are used as inputs to the OT problem, using pairwise distance as cost. By aligning data in this invariant feature space, BOT computes transport plans that respect rotational symmetries, removing nuisance variation not affecting class or content while preserving the structural relationships needed for meaningful correspondences.

4. Experiments

Understanding the Geometry of Bispectral Feature Space We conduct preliminary experiments on rotated MNIST (Deng, 2012) to visualize the geometry of bispectral features. Figure 1 shows 2D MDS embeddings of raw pixel and bispectral representations. While pixel space scatters rotated digits uniformly, the rotation-invariant bispectral space clusters them by label. Moreover, rotationally symmetric digits (e.g., 6 and 9) are brought closer in bispectral space, and the digit with the most rotational symmetries (0) yields the most tightly-packed cluster, while others with fewer symmetries (e.g., 2) are more spread out (Subfigure 1(a)). Appendix E details configurations and additional distance statistics for raw and bispectral features.

Evaluating Bispectral OT We evaluate the performance of Bispectral OT using four benchmarks for classification: the MNIST, KUZUSHIJI-MNIST (KMNIST), FASHION-MNIST (FMNIST),



(*a*) 1 Digit per Class, rotated by $\mathbb{Z}/40\mathbb{Z}$ (*b*) 10 Digits per Class, rotated by $\mathbb{Z}/40\mathbb{Z}$

Figure 1: MDS visualization of raw and $\mathbb{Z}/40\mathbb{Z}$-bispectral features for rotated MNIST digits

and EMNIST datasets (Deng, 2012; Clanuwat et al., 2018; Xiao et al., 2017; Cohen et al., 2017). Details about the datasets and experimental setup are included in Appendix D and F, respectively. We split each dataset in half, applying rotations sampled uniformly at random to the first half of images and leaving the second half unrotated. We then compute the optimal transport plan using naive OT and BOT from the rotated images to the unrotated images using various pairwise distances in the latent embedding space as the ground metric for OT. To convert the OT plan into a mapping, we assign each source image to the target class receiving the largest transported mass under its row of the coupling Γ: $\hat{y}_i = \arg\max_{k\in\{1,\dots,K\}}(\Gamma H)_{i,k}$, where H is the one-hot encoding of the labels. Table 1 measures the fraction of images that are mapped to a image of the same class in each dataset under each coupling, showing the greatly improved ability of BOT to preserve semantic label structure on datasets transformed with rotation. Statistics for raw OT using all metrics and more detailed confusion matrices with per-class matching statistics are in Appendix F.

Table 1: Class preservation accuracies for raw pixel OT vs. Bispectral OT

Dataset	**Matching Unrotated to Rotated**					**Baseline (Unrotated to Unrotated)**				
	OT	**Bispectral OT**				**OT**	**Bispectral OT**			
	(L_1)	L_1	L_2	L_2^2	cos	(L_1)	L_1	L_2	L_2^2	cos
MNIST	0.3297	0.8405	0.8020	0.8155	0.8155	0.9725	0.8603	0.8242	0.8329	0.8329
KMNIST	0.2420	0.7815	0.7225	0.7354	0.7354	0.9724	0.8143	0.7468	0.7597	0.7597
FMNIST	0.3003	0.7617	0.7662	0.7319	0.7319	0.8726	0.7982	0.7913	0.7576	0.7576
EMNIST	0.1969	0.5983	0.5693	0.5693	0.5716	0.8754	0.6416	0.6032	0.6032	0.6054

5. Discussion

In this work, we propose Bispectral OT, a symmetry-aware extension of optimal transport that computes distribution-wise distances and correspondences from bispectral representations of data invariant under group actions. Across benchmarks, we show BOT preserves semantic structure while discarding variability from transformations such as rotations. While encouraging, open challenges remain about scaling to complex non-gridded structures and richer groups or measuring distances between bispectral representations that respect their complex algebraic structure beyond common norms (L_1, L_2, cos) without losing tractability. Overall, Bispectral OT provides a promising direction for symmetry-aware distribution alignment through OT, with potential applications in transfer learning and dataset comparison in symmetry-rich domains, opening the door to a family of transport methods leveraging algebraic structure to improve robustness and interpretability of dataset comparisons.

Reproducibility Statement

The code for all experiments can be found at https://anonymous.4open.science/r/bispectral-ot-8D12/, with the specific configurations (random seeds and hyperparameters) detailed in the appendices.

References

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

David Alvarez-Melis, Tommi Jaakkola, and Stefanie Jegelka. Structured optimal transport. In *International conference on artificial intelligence and statistics*, pages 1771–1780. PMLR, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

Ching-Yao Chuang, Youssef Mroueh, Kristjan Greenewald, Antonio Torralba, and Stefanie Jegelka. Measuring generalization with optimal transport. *Advances in neural information processing systems*, 34:8294–8306, 2021.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature, 2018.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: an extension of MNIST to handwritten letters. *CoRR*, abs/1702.05373, 2017.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain Adaptation with Regularized Optimal Transport. In *Machine Learning and Knowledge Discovery in Databases*, pages 274–289, 2014.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Charlie Frogner, Farzaneh Mirzazadeh, and Justin Solomon. Learning Embeddings into Entropic Wasserstein Spaces. In *International Conference on Learning Representations*, 2019.

Ramakrishna Kakarala. A group-theoretic approach to the triple correlation. In *[1993 Proceedings] IEEE Signal Processing Workshop on Higher-Order Statistics*, pages 28–32. IEEE, 1993.

Ramakrishna Kakarala. Completeness of bispectrum on compact groups. *arXiv preprint arXiv:0902.0196*, 1, 2009.

Ramakrishna Kakarala. The bispectrum as a source of phase-sensitive invariants for fourier descriptors: a group-theoretic approach. *Journal of Mathematical Imaging and Vision*, 44:341–353, 2012.

Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

Imre Risi Kondor. *Group theoretical methods in machine learning*. Columbia University, 2008.

Risi Kondor. A novel set of rotationally and translationally invariant features for images based on the non-commutative bispectrum. *arXiv preprint cs/0701127*, 2007.

Yicheng Luo, Zhengyao Jiang, Samuel Cohen, Edward Grefenstette, and Marc Peter Deisenroth. Optimal transport for offline imitation learning. *arXiv preprint arXiv:2303.13971*, 2023.

Simon Mataigne, Johan Mathe, Sophia Sanborn, Christopher Hillar, and Nina Miolane. The selective g-bispectrum and its inversion: Applications to g-invariant networks. *Advances in Neural Information Processing Systems*, 37:115682–115711, 2024.

Boris Muzellec and Marco Cuturi. Generalizing point embeddings using the wasserstein space of elliptical distributions. *Advances in Neural Information Processing Systems*, 31, 2018.

Walter Rudin. *Fourier analysis on groups*. Courier Dover Publications, 2017.

Sophia Sanborn, Christian Shewmake, Bruno Olshausen, and Christopher Hillar. Bispectral neural networks. *arXiv preprint arXiv:2209.03416*, 2022.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Appendix A. Extended Related Works

Representing collections of objects as empirical measures and comparing them via OT is an active area of research, with costs typically defined directly from features or via latent embeddings. For example, Muzellec and Cuturi (2018) models objects as elliptical distributions, and Frogner et al. (2019) represents the embeddings as discrete measures. In supervised settings, label information can be injected into the cost (Courty et al., 2014; Alvarez-Melis et al., 2018), and hierarchical formulations address discrete labels (Alvarez-Melis and Fusi, 2020). We seek to build upon this line of work by using a bispectral representation of an object to encode symmetries within a dataset.

In another line of work, the theory of the group-invariant bispectrum was primarily developed in Kakarala (1993, 2009, 2012) for signal processing contexts. The invariant bispectrum was first introduced to machine learning in Kondor (2007, 2008), utilizing embeddings of the non-commutative bispectrum in vision tasks. More recently, Sanborn et al. (2022) described a neural network architecture using the bispectrum to learn groups from the data, and Mataigne et al. (2024) introduced an algorithm to reduce the computational cost of the group bispectrum and other similar invariants. To the authors knowledge, this is the first work that encodes symmetry awareness into OT, bridging these two lines of work.

Appendix B. Background on Group Representation Theory

We introduce the fundamentals of group representation theory, which serves as the foundation of the theory of the group bispectrum.

Definition 1 (Group) *A* group $(G, \cdot)$ *is a set* G *with a binary operation referred to as the group product that satisfies the following axioms:*

1. Closure: *For all* $a, b \in G$, *we have* $ab \in G$.
2. Associativity: *For any* $a, b, c \in G$, *we have that* $(ab)c = a(bc)$.
3. Identity: *There exists some* identity *element* e *such that for all* $g \in G$, *we have* $eg = ge = g$.
4. Inverse: *For every element* g, *there exists an* inverse *element* g^{-1} *such that* $gg^{-1} = g^{-1}g = e$.

Concretely, a group G can define a class of transformations like rotations or translations in the plane, with each element of the group defining a particular transformation. Groups that are important to us in this paper include the planar rotation group $SO(2)$, also known as the special orthogonal group, and its discrete analog, the cyclic group $\mathbb{Z}/n\mathbb{Z} = \{0, 1, \ldots, n-1\}$ with group product addition modulo n, which is the group of all rotational symmetries of a regular n-gon. These groups are *commutative* or *abelian,* which means that the order of operations does not matter: for all $a, b \in G$ for G commutative, we have $ab = ba$. This is in contrast to *non-commutative* groups G, where there exists some $a, b \in G$ such that

$ab \neq ba$. Examples of non-commutative symmetry group include $O(2)$, the group of all planar rotations and reflections, $SO(3)$, the group of all 3D rotations, and D_n, the group of symmetries of an n-gon, which can be viewed as the discrete version of $O(2)$.

Definition 2 (Group Homomorphism) *A* group homomorphism *between two groups G and H with operations $\cdot$ and $*$, respectively, is a map $\rho : G \to H$ that respects the underlying group structure of G and H, i.e. $\rho(u \cdot v) = \rho(u) * \rho(v)$. If such a map exists, then G and H are called* homomorphic. *If such a map is a bijection, it is then called an* isomorphism.

Note that isomorphic groups are essentially the same group, but arising in different contexts. This motivates *representation theory*, which studies groups via linear actions on vector spaces, using linear algebra to make abstract structures concrete.

Definition 3 (Group Representation) *A* representation *of a group G is a group homomorphism $\rho : G \to GL(V)$ assiging elements of G to elements of the group of linear transformations over a vector space V. In most contexts, V is $\mathbb{R}^n$ or $\mathbb{C}^n$.*

A representation is *reducible* if there exists a change of basis that decomposes the representation into a direct sum of other representations. An *irreducible representation* cannot be decomposed in this way, and the set of them $Irr(G)$ are often called the *irreps* of G. For all finite groups and compact groups, which is the only classes of groups we will consider, the irreps consist only of unitary transformations, so throughout this paper, we assume $\rho(g^{-1}) = \rho(g)^{-1} = \rho(g)^\dagger$ for $\dagger$ the conjugate transpose. For commutative groups, the irreducible representations are scalars and in bijection with the group elements. This nice characterization does not extend to non-commutative groups, where the irreducible representations are matrix valued of variable dimension.

The discussion of using symmetry groups to explain natural transformations in datasets motivates the definition of a *group action,* which we provide now:

Definition 4 (Group Action) *For G a group and X a set, a* group action *is a map $T : G \times X \to X$ with the following properties:*

1. *The identity e maps an element $x \in X$ to itself: for all $x \in X$,, we have $T(e, x) = x$.*
2. *For all $g_1, g_2 \in G$, we have $T(g_1, T(g_2, x)) = T(g_1 g_2, x)$.*

For simplicity, given a group action T, we often say that a point x maps to gx $(= T(g, x))$. Then, if X is a space on which a group G acts, we define the *orbit* of a point $x \in X$ to be the set $\{gx : g \in G\}$. In the context of image transformations, the orbit is the set of all transformed versions of an image. For example, if $G = SO(2)$, the orbit contains all rotated versions of that image. Using group actions, we can also define the concepts of *invariance* and *equivariance,* which are critical to literature in geometric machine learning.

Definition 5 (Invariance) *For sets X, Y, a function $f : X \mapsto Y$ is* G-invariant *if $f(x) = f(gx)$ for all $g \in G$ and $x \in X$. In other words, group actions on the input space have no effect on the output.*

Definition 6 (Equivariance) *For sets X, Y, a function $f : X \mapsto Y$ is* G-equivariant *if $f(gx) = g'f(x)$ for all $g \in G$ and $x \in X$, where $g' \in G'$, a group homomorphic to G. In other words, group actions on the input space results in a corresponding group action on the output space.*

The bispectrum of the group G is an example of a G-invariant function, which we exploit in the representations of images used in the paper.

Appendix C. The Non-Commutative Bispectrum

The bispectrum has an analogous form in the setting of non-commutative groups, but is adjusted to account for the fact that the irreducible representations of a non-commutative group are not necessarily one dimensional, unlike in the compact commutative case. This more general form of the bispectrum is defined to be

$$\beta_{\rho_i,\rho_j} = [\hat{f}_{\rho_i} \otimes \hat{f}_{\rho_j}] C_{\rho_i,\rho_j} \left[\bigoplus_{\rho \in \rho_i \otimes \rho_j} \hat{f}_{\rho}^{\dagger} \right] C_{\rho_i,\rho_j}^{\dagger}, \tag{5}$$

where $\oplus$ is a direct sum over irreducible representations, $\otimes$ is a tensor product, and C_{ρ_i,ρ_j} is a unitary matrix defining a Clebsch-Gordan decomposition on the tensor product of a pair of irreducible representations (Kondor, 2007).

Appendix D. Dataset Details

Information about the datasets used, including references, are provided in Table 2. MNIST has one class for each digit, KUZUSHIJI-MNIST has classes corresponding to distinct cursive Japanese calligraphy characters, FASHION-MNIST has classes corresponding to different articles of clothing, and the letters split of EMNIST has one class for each letter of the alphabet.

Table 2: Summary of all the datasets used in this work. For all experiments, we normalize the dataset to have mean 0 and standard deviation 1.

Dataset	Input Dimension	Number of Classes	Train Examples	Test Examples	Source
MNIST	28×28	10	60K	10K	(Deng, 2012)
KUZUSHIJI-MNIST	28×28	10	60K	10K	(Clanuwat et al., 2018)
FASHION-MNIST	28×28	10	60K	10K	(Xiao et al., 2017)
EMNIST (letters)	28×28	26	145K	10K	(Cohen et al., 2017)

Appendix E. $O(2)$-mnist Preliminary Experiments

E.1. Experimental Setup

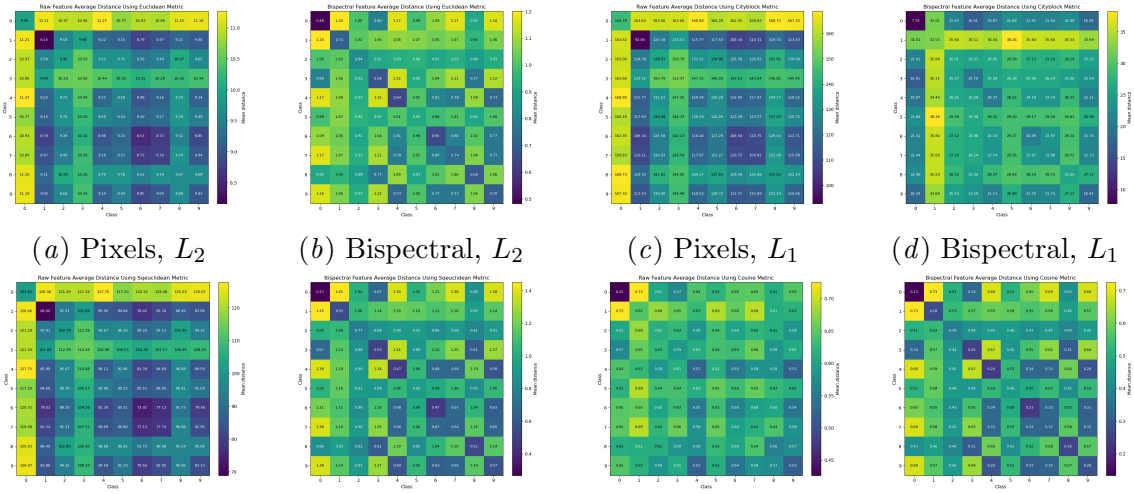
Throughout this paper, we use $R = \frac{\min(M,N)}{2}$ radial bins and $K = 40$ angular bins in the discretized polar representation of our images of size $M \times N$ for the sake of computational feasibility. Thus, the bispectrum we compute is the $\mathbb{Z}/40\mathbb{Z}$ bispectrum, which is invariant to group actions that define rotations by $\frac{360^\circ}{40} = 9^\circ$. Due to the discrete grid structure of

features in pixel space, increasing K does not necessarily give us better symmetry-invariant representations – our discrete polar representation is of size $R \times K$, and reflects at most $M \times N$ pixels of information. Since $R \sim \min(M, N)$, the number of angular bins must be on the order of $K \sim \max(M, N)$, which in the case of MNIST, is $M = N = 28$.

To construct the 2D MDS embeddings of the pixel and bispectral representations, we randomly sampled representative images from each class of the train split of MNIST, and rotated each by all multiples of 9° from 0° to 360°. Specifically, Figure 1(a) depicts MDS embeddings of the pixel and bispectral representations of rotations of one representative image for clearer visualization of cluster structure (for a total of $10 \cdot 40 = 400$ embedded features), and Figure 1(b) depicts MDS embeddings of the pixel and bispectral representations of rotations of ten representative images per class for a more global visualization of the embeddings with intra-class variation (for a total of $10 \cdot 10 \cdot 40 = 4000$ embedded features). The code to generate these plots is included in the linked repository (seed 8).

E.2. Additional Inter-Class Distance Statistics

We also include the following confusion matrices in Figure 2 depicting the average inter-class and intra-class distance statistics for the raw and bispectral representations of ten randomly sampled images of each class of MNIST rotated by multiples of 9° using different geometric distances. Euclidean distance denotes the L_2 norm $\sqrt{\sum_i (x_i - y_i)^2}$, cityblock distance denotes the L_1 norm $\sum_i |x_i - y_i|$, sqeuclidean denotes the squared L_2 norm $\sum_i (x_i - y_i)^2$, and cosine denotes the cosine distance, defined as $1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2}$ between two vectors where $\cdot$ denotes the dot product and $\| * \|_2$ is the L_2 norm.



(a) Pixels, L_2 (b) Bispectral, L_2 (c) Pixels, L_1 (d) Bispectral, L_1

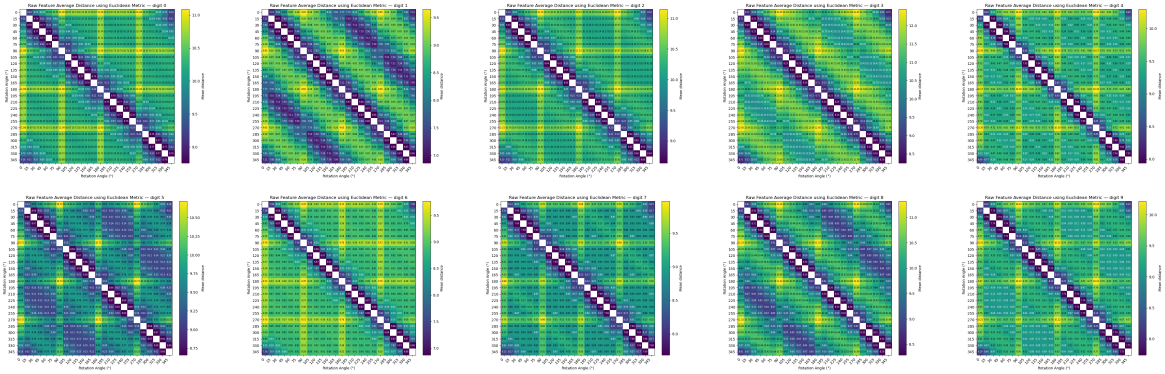
(e) Pixels, L_2^2 (f) Bispectral, L_2^2 (g) Pixels, cos (h) Bispectral, cos

Figure 2: Average inter-class and intra-class distances for pixel and bispectral representations of MNIST digits using different metrics.

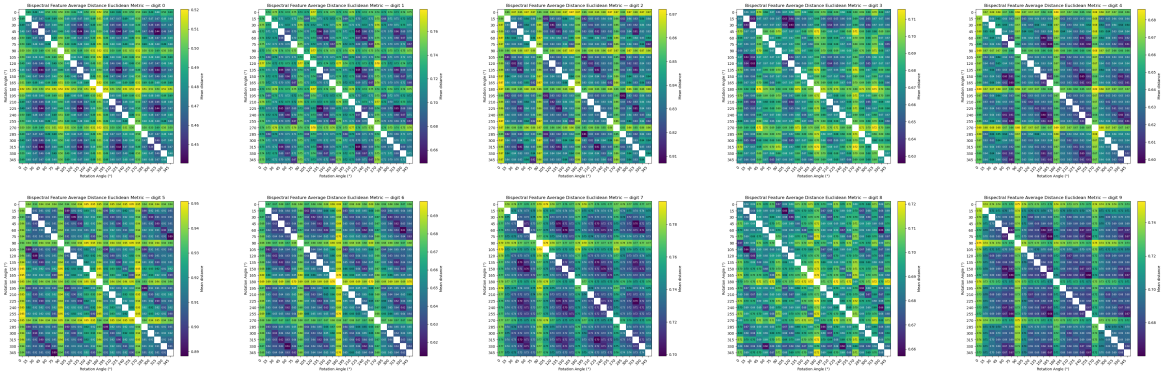
E.3. Additional Per-Digit Distance Statistics

In addition to the confusion matrices in Figure 2, we include the average distance between images of the same class that are rotated by different angles for a better understanding of the local geometry within each class of the bispectral features, in comparison to the raw pixel representations of images. For greater ease of visualization, we rotate each of our 10 sampled images from each class by each multiple of 15° from 0° to 360°, and continue to use the $\mathbb{Z}/40\mathbb{Z}$-bispectrum in the bispectral representation of images. The grid-like patterns are likely a function of how we handled clipping due to rotations.

The distance statistics using the L_2 distance are depicted in Figure 3, the distance statistics using the L_1 distance are depicted in Figure 4, the distance statistics using the L_2^2 distance are depicted in Figure 5, and the distance statistics using cosine similarity are depicted in Figure 6.

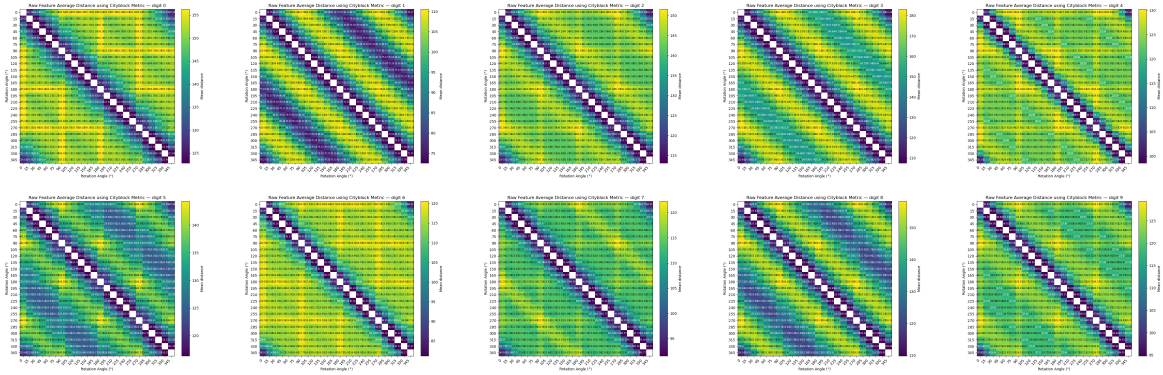


(*a*) Raw feature representations

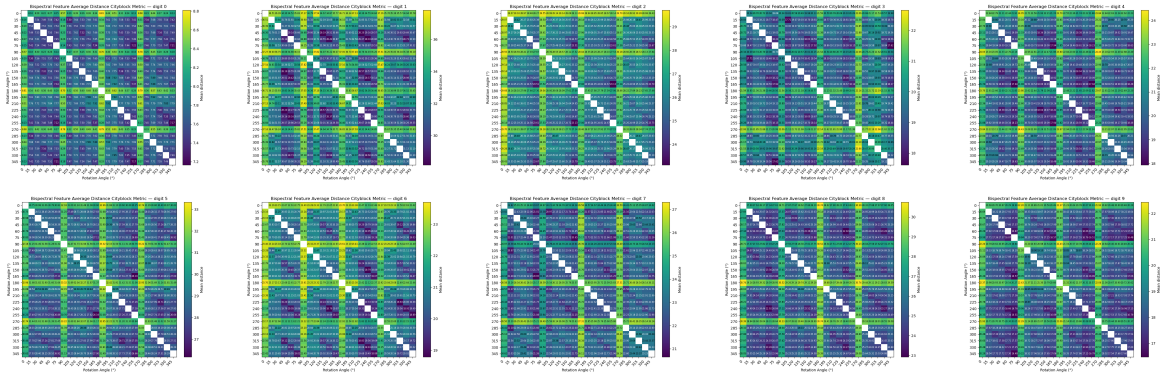


(*b*) Bispectral feature representations

Figure 3: Average distance between bispectral and pixel representations of rotated images from the same class of MNIST using euclidean norm.

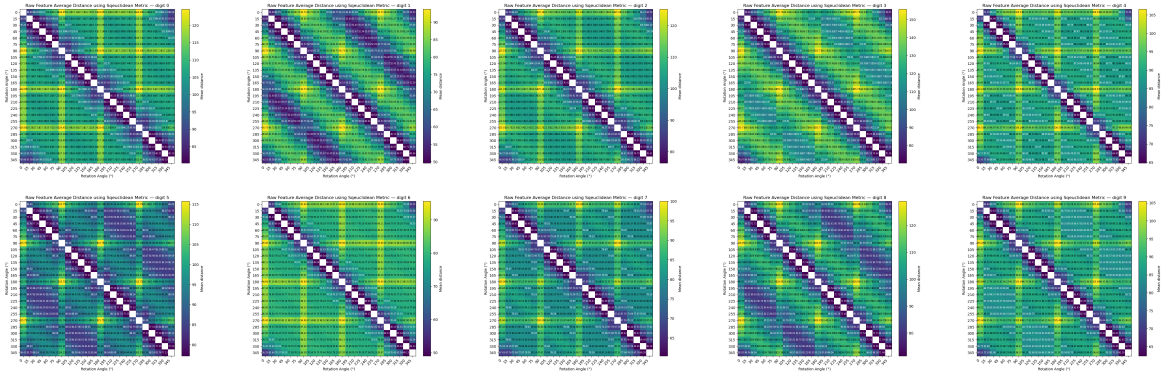


(a) Raw feature representations

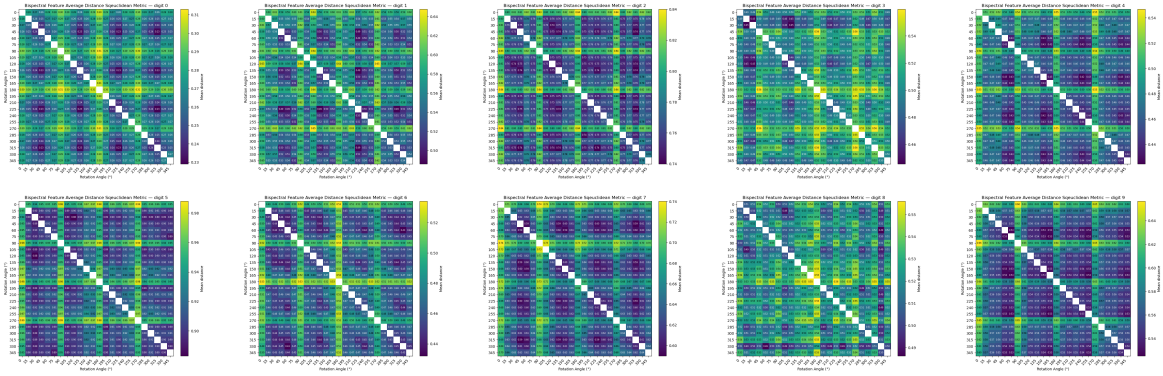


(b) Bispectral feature representations

Figure 4: Average distance between bispectral and pixel representations of rotated MNIST images using L_1 norm.

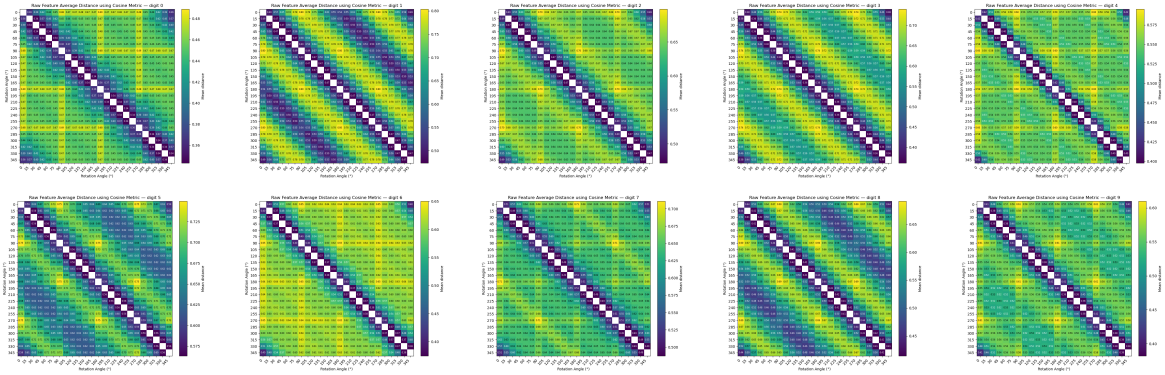


(a) Raw feature representations

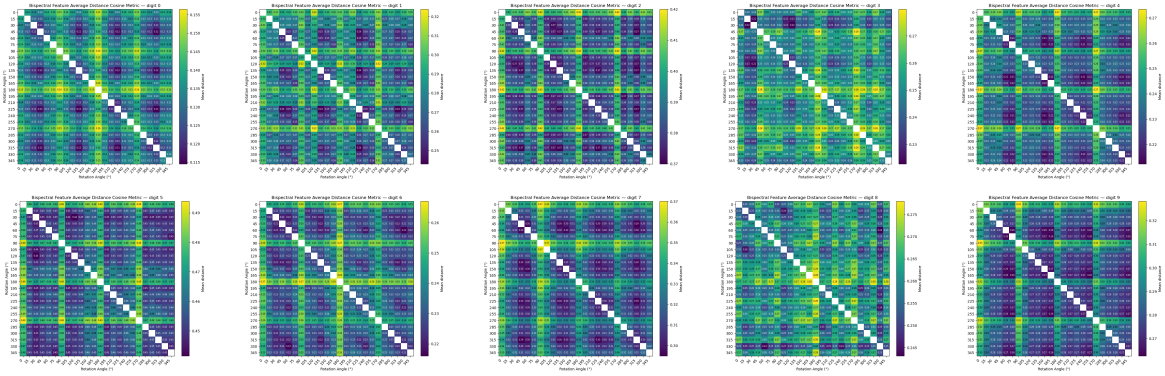


(b) Bispectral feature representations

Figure 5: Average distance between bispectral and pixel representations of rotated MNIST images using squared euclidean distance.



(a) Raw feature representations



(b) Bispectral feature representations

Figure 6: Average distance between bispectral and pixel representations of rotated MNIST images using cosine similarity.

Appendix F. Benchmark Dataset Experiments

F.1. Experimental Setup

As described in Appendix E for the preliminary experiments, we use $R = \frac{\min(M,N)}{2}$ radial bins and $K = 40$ angular bins in the discretized polar representation of our images of size $M \times N$. Thus, the bispectral representations of our features are $\mathbb{Z}/40\mathbb{Z}$ rotation invariant, and constructed using the group bispectrum of $\mathbb{Z}/40\mathbb{Z}$.

We perform OT between two disjoint halves of the `torchvision`'s[1] training split of each of MNIST, KMNIST, FASHION-MNIST, and the letters split of EMNIST, randomly splitting (with seed 0) each class in each dataset's training set into two disjoint halves to construct the two datasets that we perform OT between. For KMNIST and FASHION-MNIST, the training split is perfectly balanced between classes, so we obtain two datasets of size 30,000, with exactly 3,000 images per class. The MNIST training split is only approximately balanced, so splitting each class in half yields two datasets of size 29,997, with class sizes distributed as follows: $\{0:2961, 1:3371, 2:2979, 3:3065, 4:2921, 5:2710, 6:2959, 7:3132, 8:2925, 9:2974\}$. For EMNIST (letters), which contains 26 classes with a total of 124,800 training examples, we similarly split each class in half, yielding two datasets of size 62,400 with exactly 2,400 images per class.

For the main experiment, we perform OT from a set of images with rotations sampled uniformly at random from 0° to 360° to a disjoint unrotated set. As a baseline, also compute the class preservation statistics for the transport plan computed on two sets of unrotated raw and bispectral features. We calculate OT using `greenkhorn` greedy implementation of the entropically regularized version of OT in the Sinkhorn implementation of the `pot`[2] package (Flamary et al., 2021; Altschuler et al., 2017) for computational feasibility using a regularization of 0.01 (which maximizes class preservation accuracy while ensuring convergene). Table 1 includes only the class preservation accuracies for raw OT using the L_1 pairwise cost between images (since this performs the best overall), so for completeness, we include the statistics for the other ground metrics in Table 3.

Table 3: Class preservation accuracies for naive OT using L_1, L_2, L_2^2 and cosine pairwise metrics for cost.

	Unrotated to Rotated				**Baseline**			
Dataset	L_1	L_2	L_2^2	cos	L_1	L_2	L_2^2	cos
MNIST	0.3297	0.3322	0.3289	0.3239	0.9725	0.9697	0.9742	0.9705
KMNIST	0.2420	0.2408	0.2348	0.2412	0.9724	0.9698	0.9712	0.9666
FMNIST	0.3003	0.2787	0.2680	0.2615	0.8726	0.8748	0.8466	0.8736
EMNIST	0.1969	0.1991	0.1991	0.1966	0.8754	0.8678	0.8678	0.8730

F.2. Per-Class OT Matching Statistics

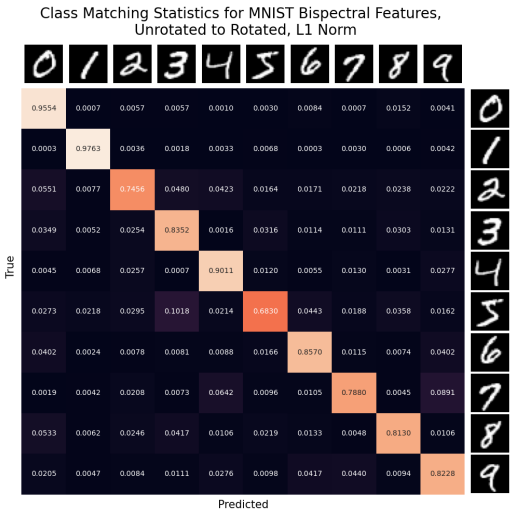
For a more granular understanding of the how BOT preserves semantic label structure in each dataset, we include confusion matrices that separate the class preservation accuracies

1. https://docs.pytorch.org/vision/stable/index.html
2. pythonot.github.io/

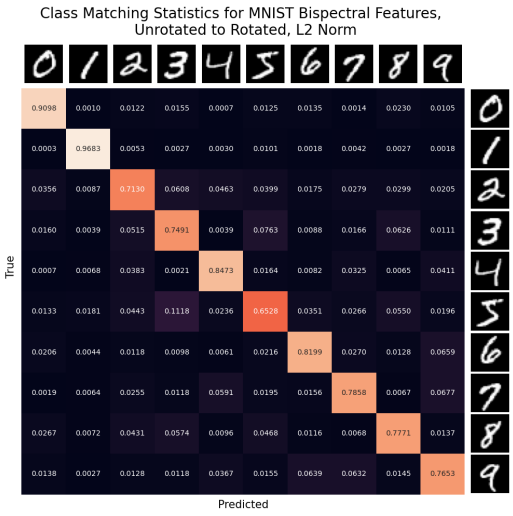
for raw feature OT and Bispectral OT in Table 1 by class, detailing the fraction of each class matched to elements in each other class for each of our benchmark datasets.

F.2.1. MNIST

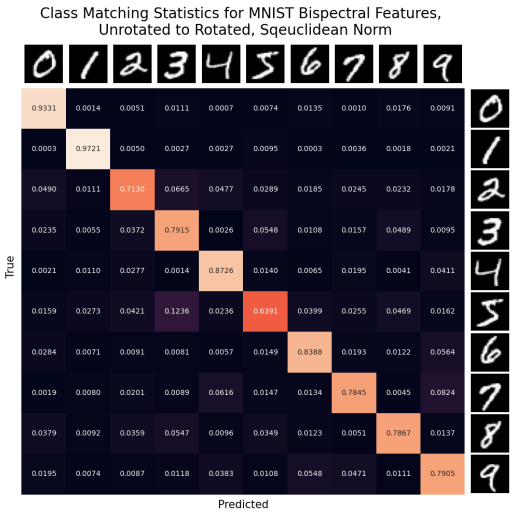
We include the confusion matrices for OT on MNIST for the OT matching between rotated and unrotated images in Figures 7 (bispectral features) and 8 (raw pixel features), and for the baseline experiment (matching unrotated images) in Figures 9 (bispectral features) and 10 (raw pixel features)



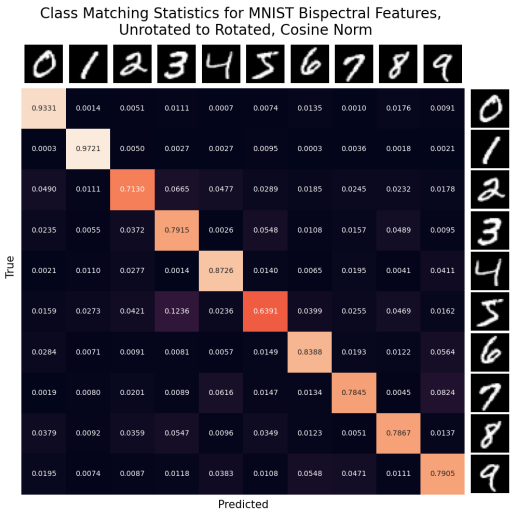
(a) L_1 Cost



(b) L_2 Cost

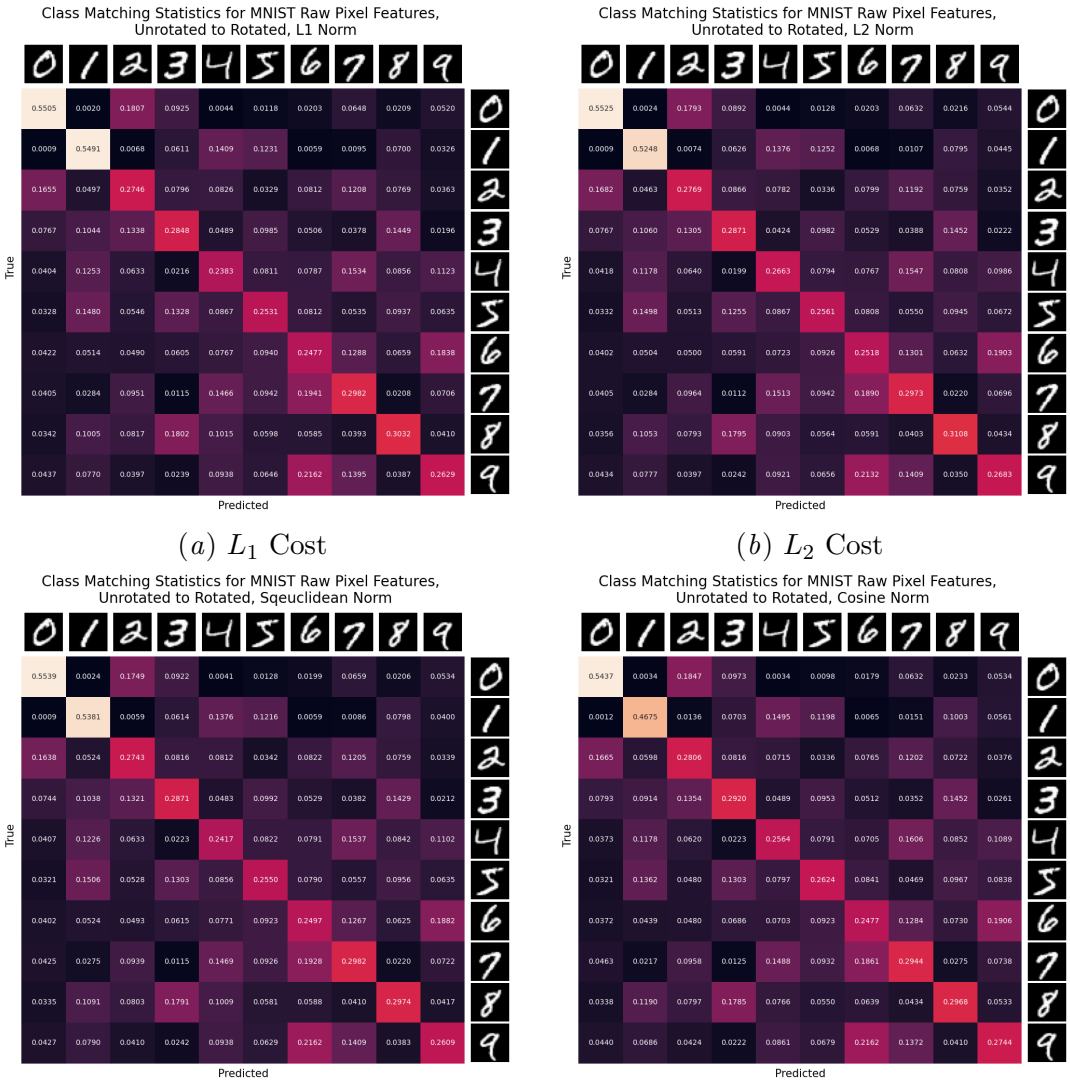


(c) L_2^2 Cost



(d) cos Cost

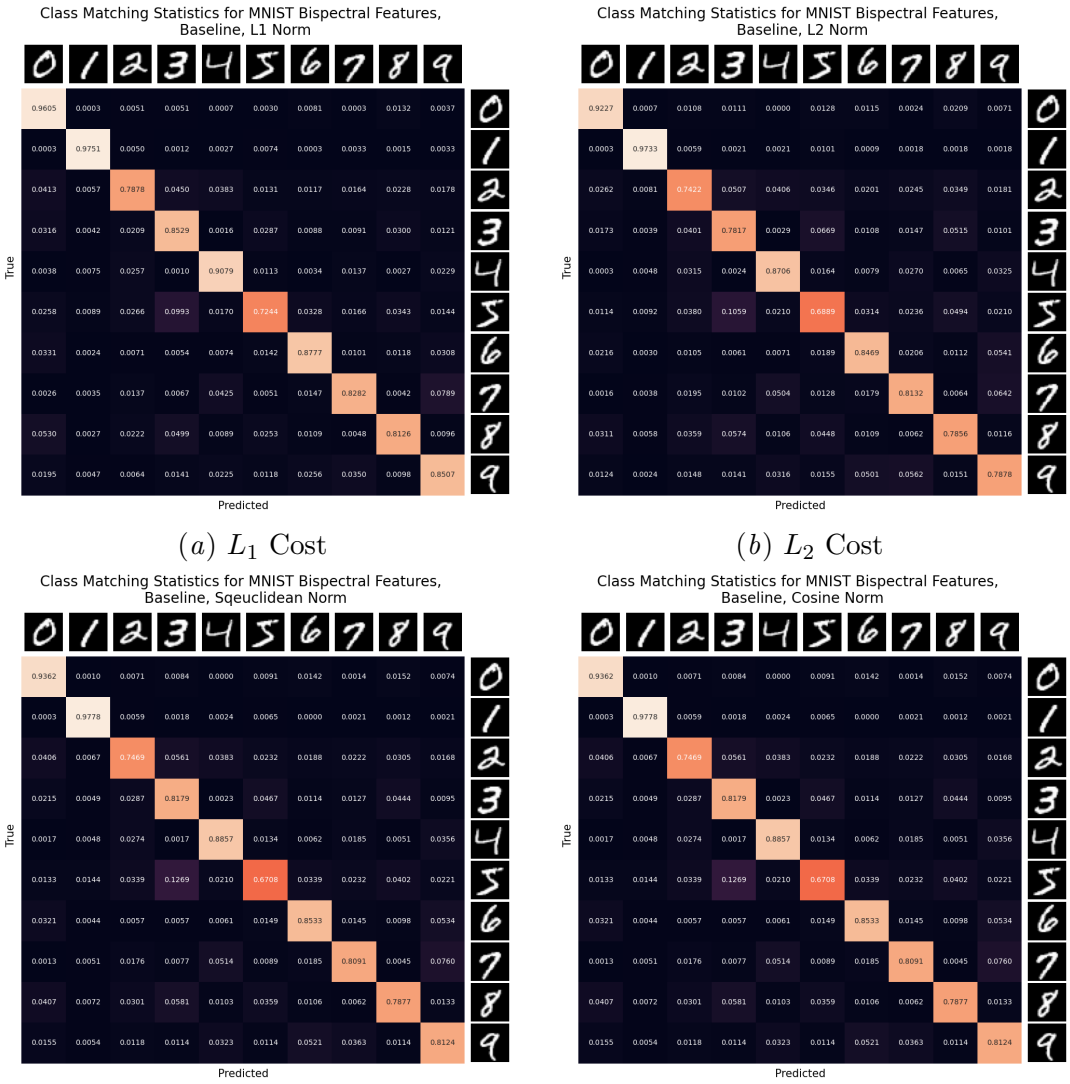
Figure 7: Class matching statistics for OT plan from rotated MNIST to unrotated MNIST on bispectral features.



(a) L_1 Cost (b) L_2 Cost

(c) L_2^2 Cost (d) cos Cost

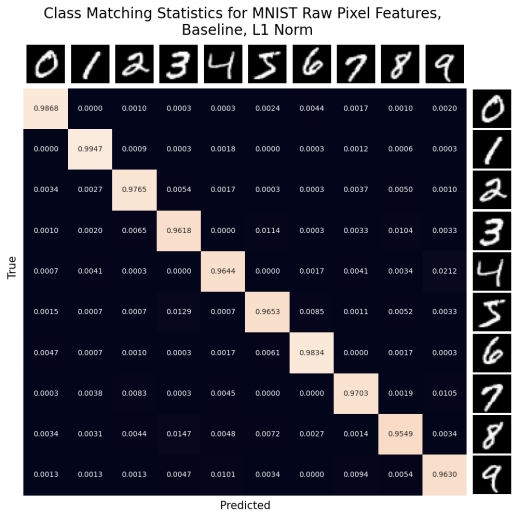
Figure 8: Class matching statistics for OT plan from rotated MNIST to unrotated MNIST on raw pixel features.



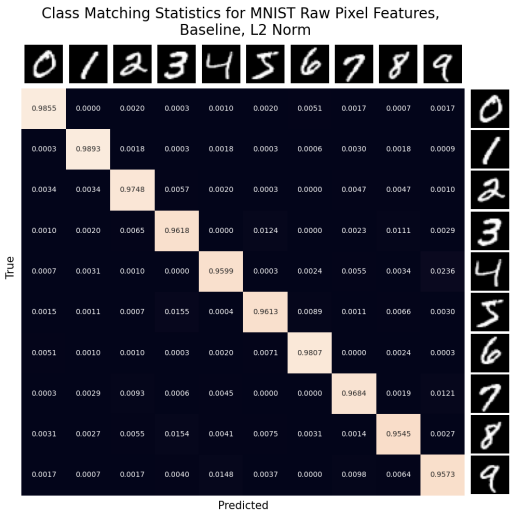
(a) L_1 Cost (b) L_2 Cost

(c) L_2^2 Cost (d) cos Cost

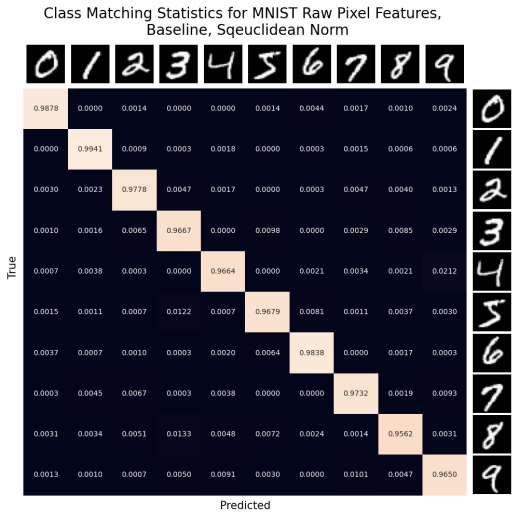
Figure 9: Class matching statistics for OT plan from unrotated MNIST to unrotated MNIST (baseline) on bispectral features.



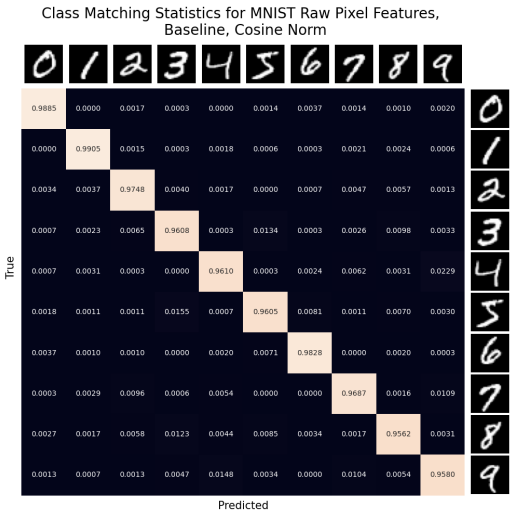
(a) L_1 Cost



(b) L_2 Cost



(c) L_2^2 Cost

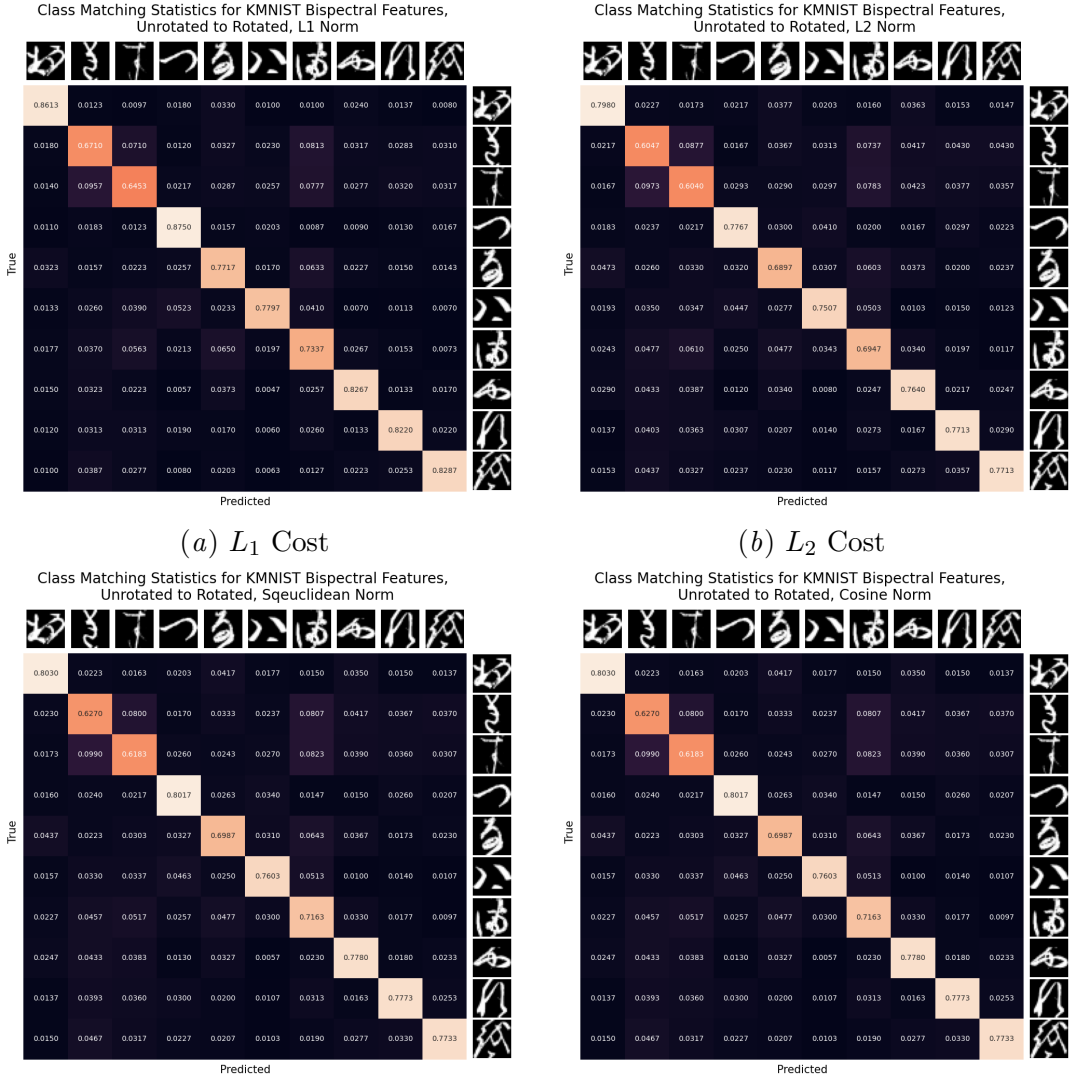


(d) cos Cost

Figure 10: Class matching statistics for OT plan from unrotated MNIST to unrotated MNIST (baseline) on raw pixel features.

F.2.2. KMNIST

We include the confusion matrices for OT on KMNIST for the OT matching between rotated and unrotated images in Figures 11 (bispectral features) and 12 (raw pixel features), and for the baseline experiment (matching unrotated images) in Figures 13 (bispectral features) and 14 (raw pixel features).



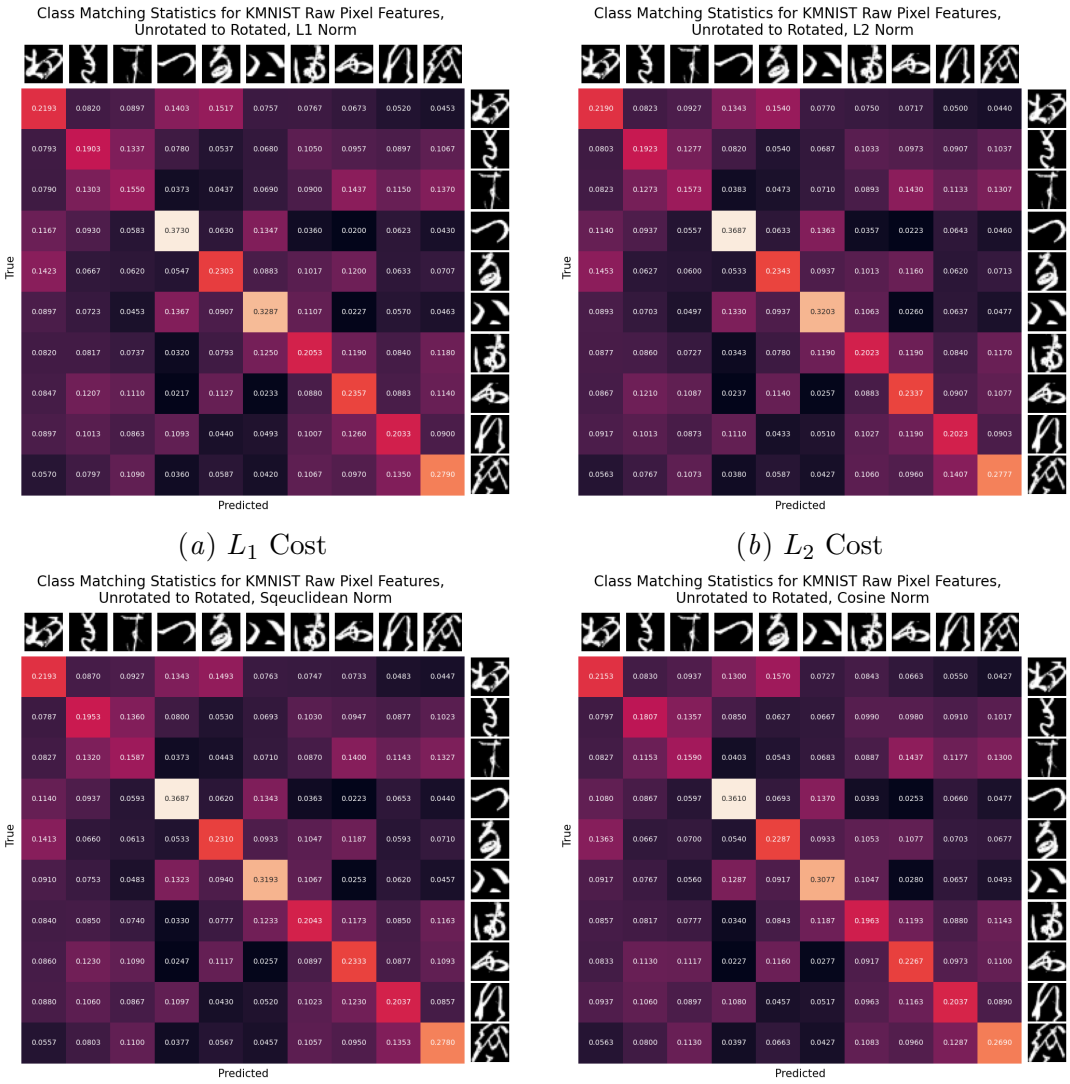
(a) L_1 Cost (b) L_2 Cost

(c) L_2^2 Cost (d) cos Cost

Figure 11: Class matching statistics for OT plan from rotated KMNIST to unrotated KMNIST on bispectral features.

F.2.3. FASHION-MNIST

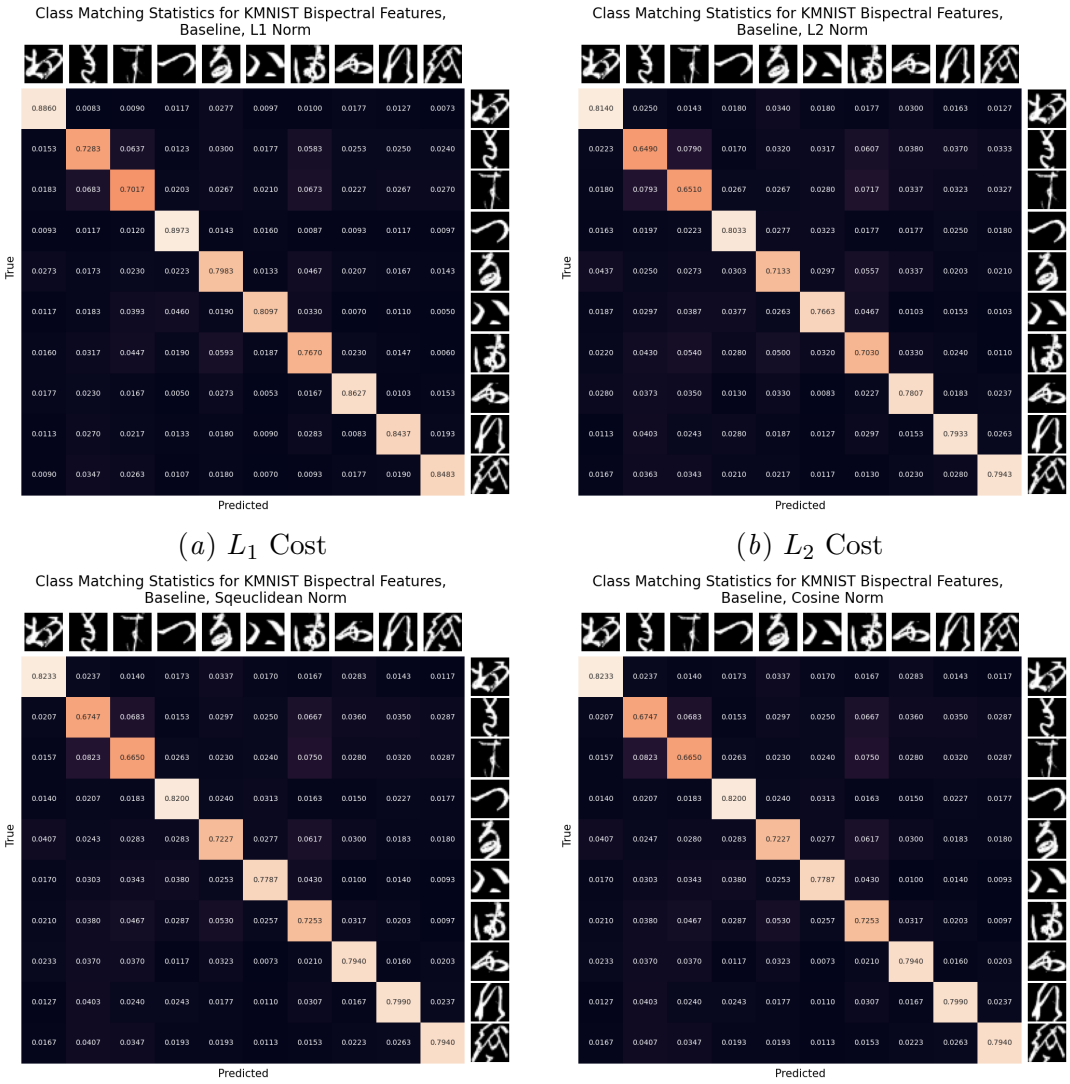
We include the confusion matrices for OT on FASHION-MNIST for the OT matching between rotated and unrotated images in Figures 15 (bispectral features) and 16 (raw pixel features),



(a) L_1 Cost (b) L_2 Cost

(c) L_2^2 Cost (d) cos Cost

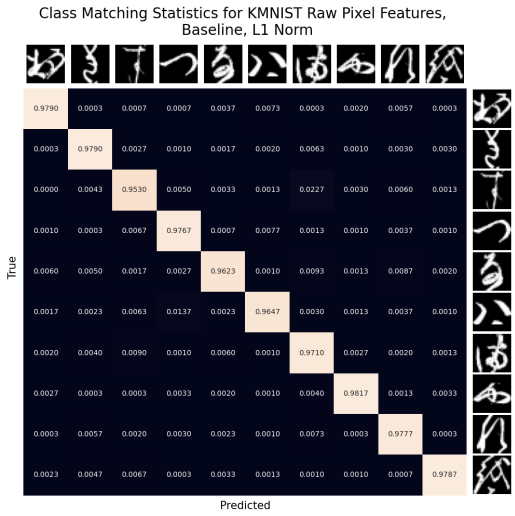
Figure 12: Class matching statistics for OT plan from rotated KMNIST to unrotated KMNIST on raw pixel features.



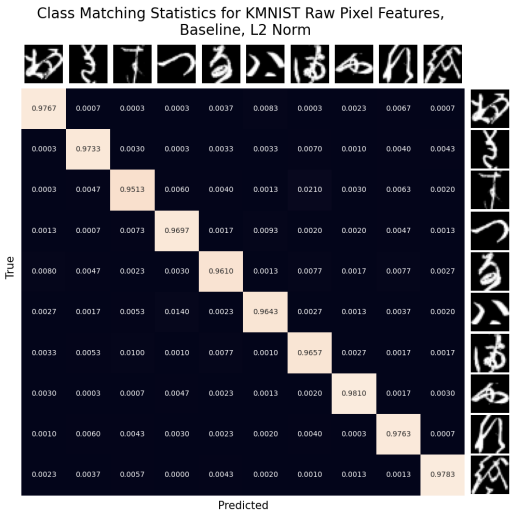
(a) L_1 Cost (b) L_2 Cost

(c) L_2^2 Cost (d) cos Cost

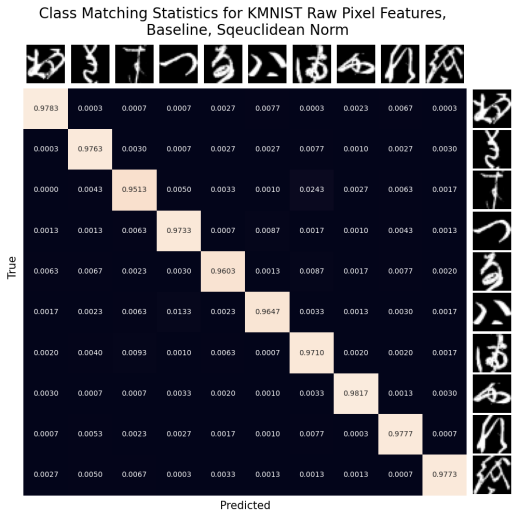
Figure 13: Class matching statistics for OT plan from unrotated KMNIST to unrotated KMNIST (baseline) on bispectral features.



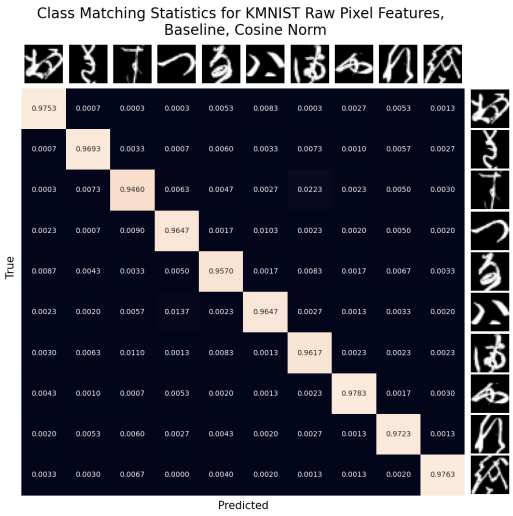
(a) L_1 Cost



(b) L_2 Cost



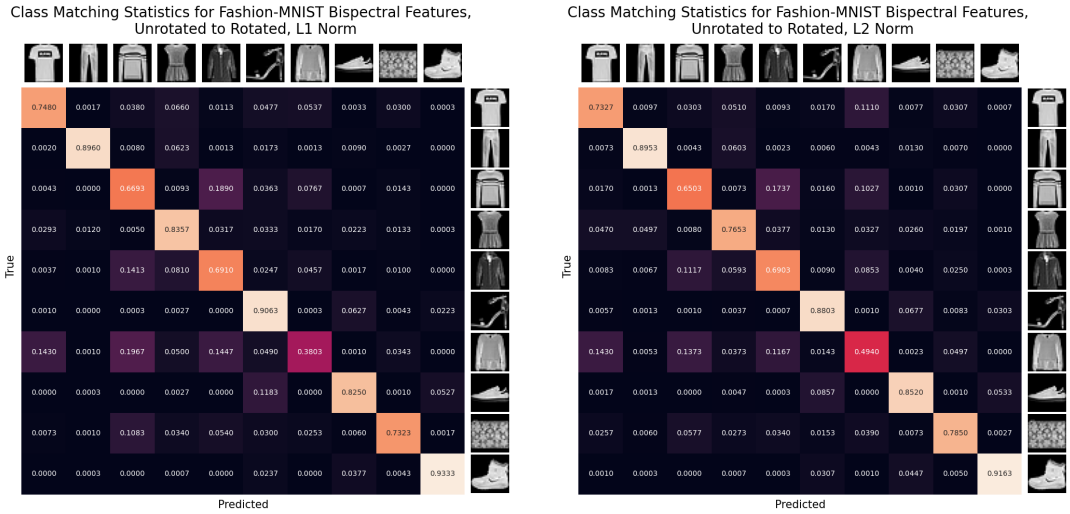
(c) L_2^2 Cost



(d) cos Cost

Figure 14: Class matching statistics for OT plan from unrotated KMNIST to unrotated KMNIST (baseline) on raw pixel features.

and for the baseline experiment (matching unrotated images) in Figures 17 (bispectral features) and 18 (raw pixel features).



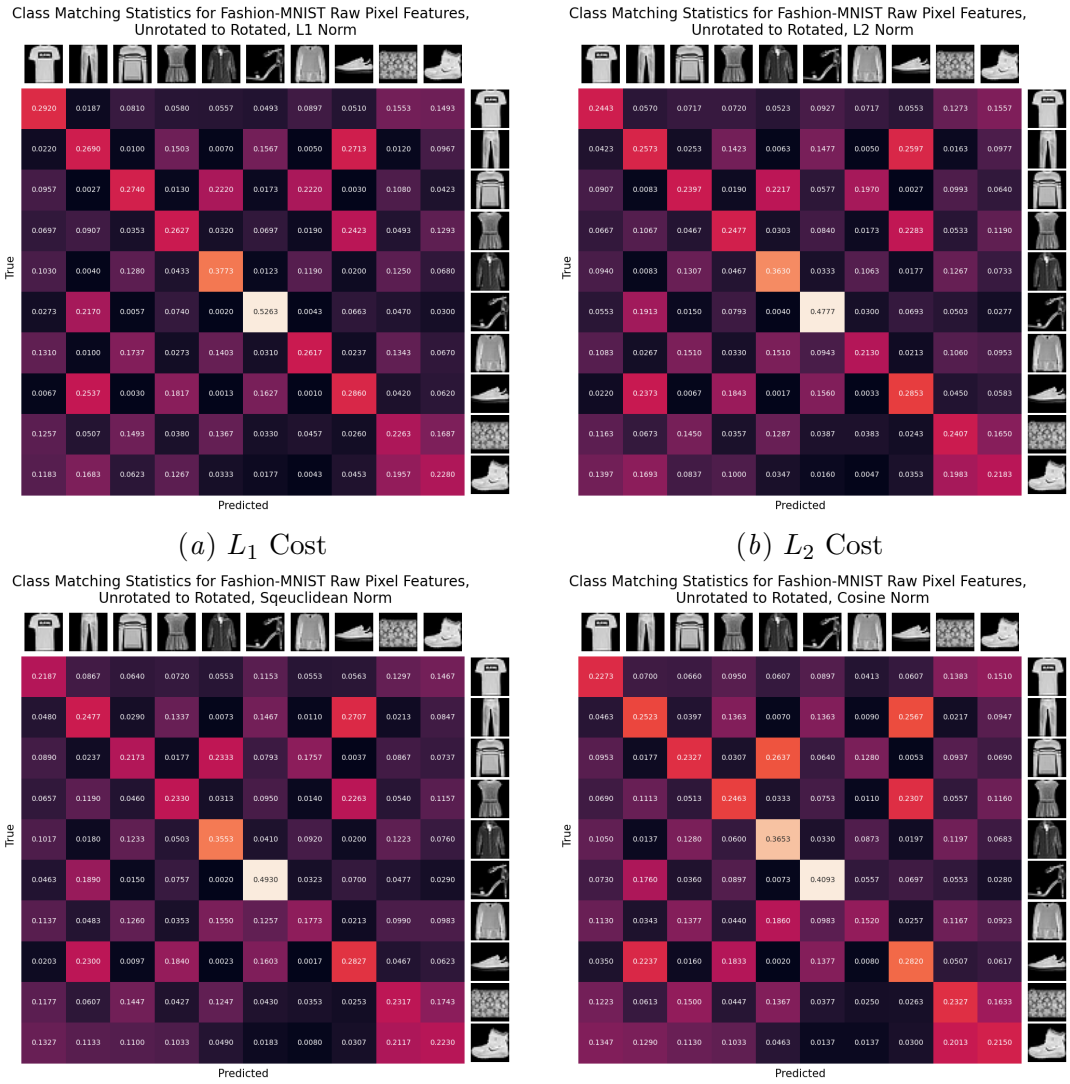
(a) L_1 Cost (b) L_2 Cost

(c) L_2^2 Cost (d) cos Cost

Figure 15: Class matching statistics for OT plan from rotated FASHION-MNIST to unrotated FASHION-MNIST on bispectral features.

F.2.4. EMNIST

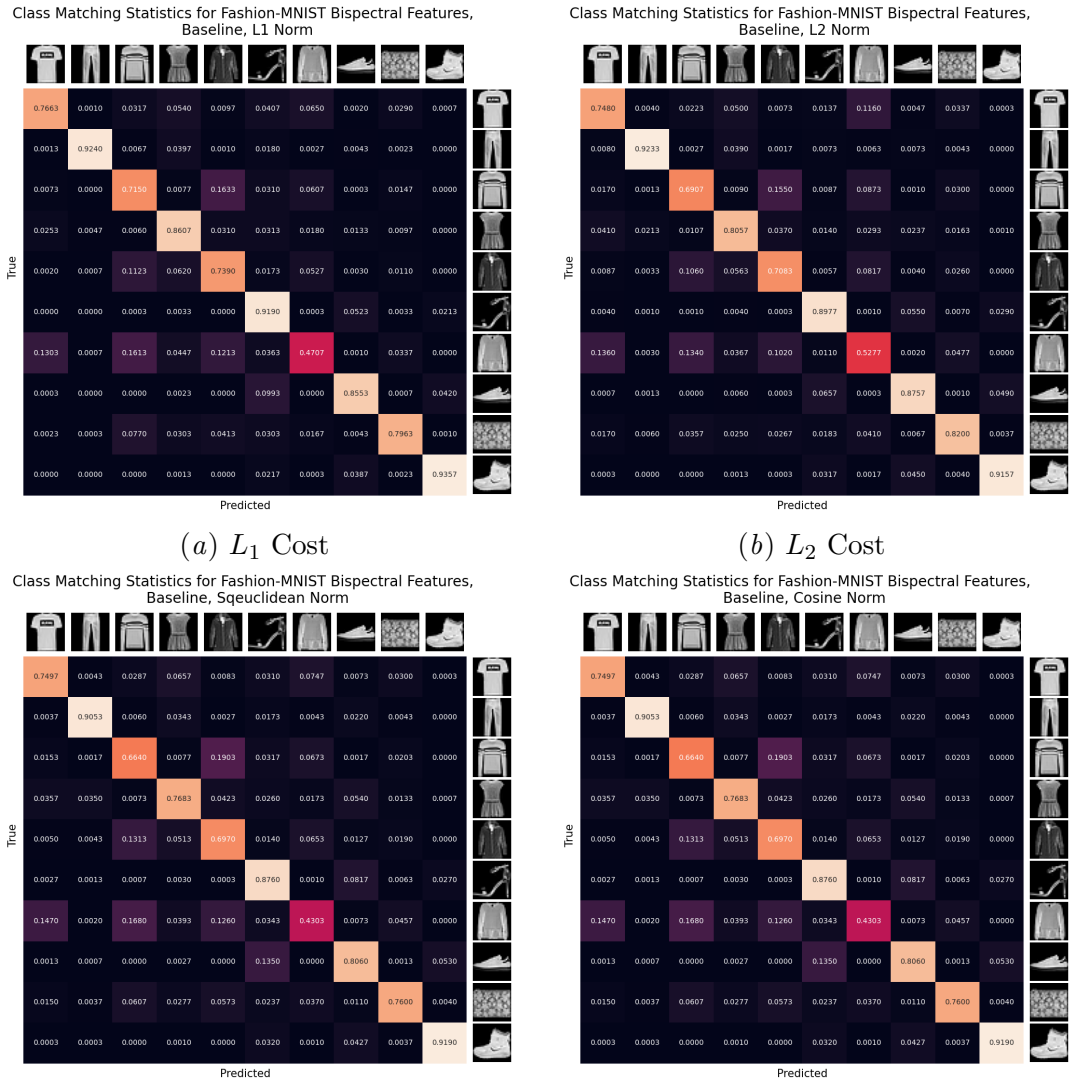
We include the confusion matrices for OT on FASHION-MNIST for the OT matching between rotated and unrotated images in Figures 19 (bispectral features) and 20 (raw pixel features), and for the baseline experiment (matching unrotated images) in Figures 21 (bispectral features) and 22 (raw pixel features).



(a) L_1 Cost (b) L_2 Cost

(c) L_2^2 Cost (d) cos Cost

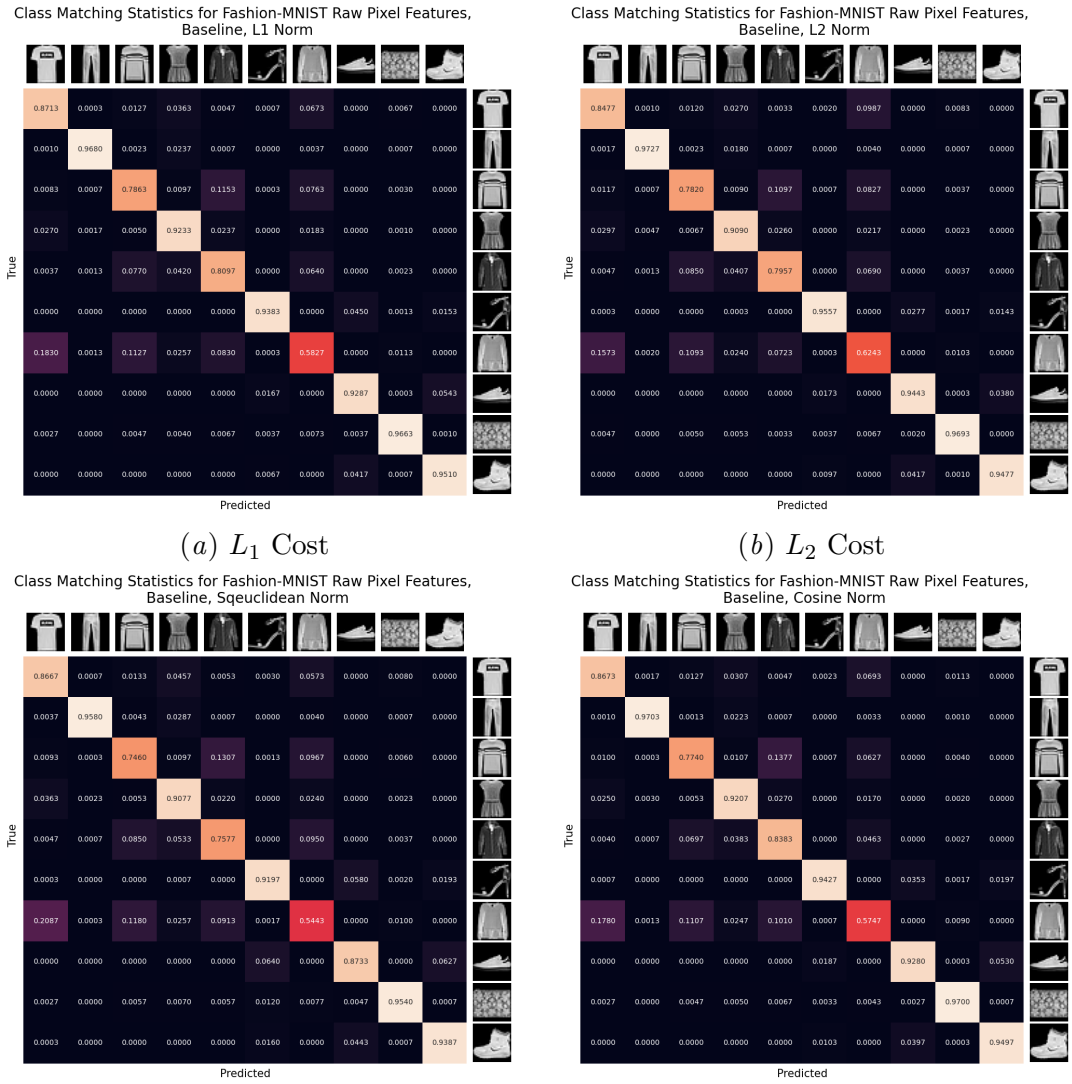
Figure 16: Class matching statistics for OT plan from rotated FASHION-MNIST to unrotated FASHION-MNIST on raw pixel features.



(a) L_1 Cost (b) L_2 Cost

(c) L_2^2 Cost (d) cos Cost

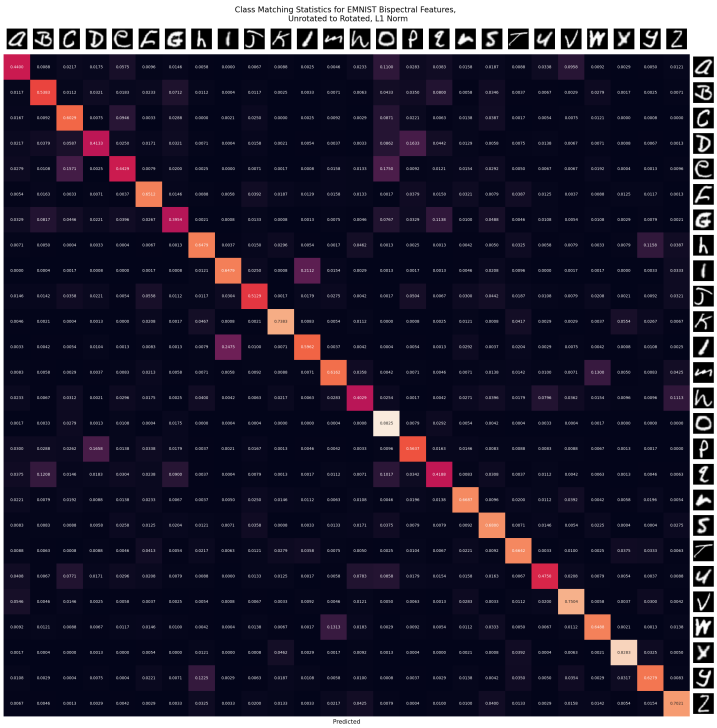
Figure 17: Class matching statistics for OT plan from unrotated FASHION-MNIST to unrotated FASHION-MNIST (baseline) on bispectral features.



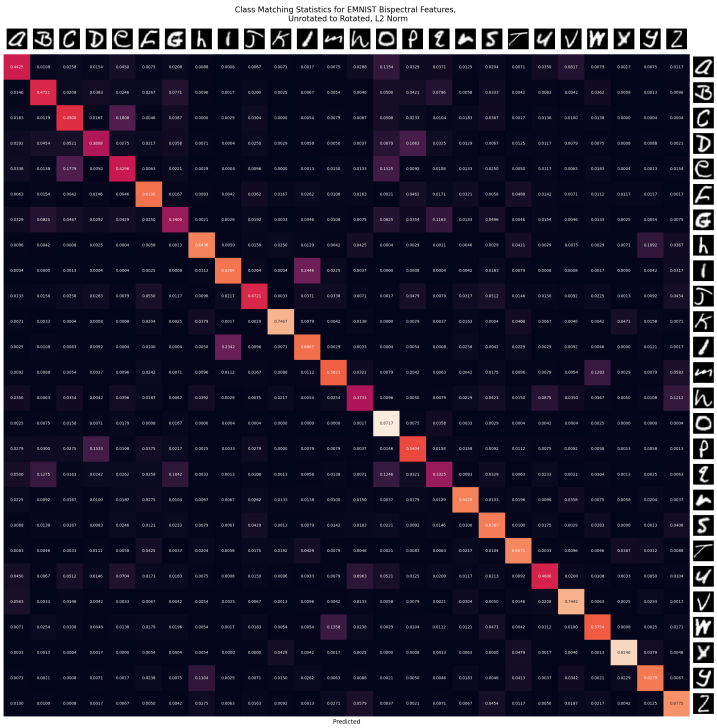
(a) L_1 Cost (b) L_2 Cost

(c) L_2^2 Cost (d) cos Cost

Figure 18: Class matching statistics for OT plan from unrotated FASHION-MNIST to unrotated FASHION-MNIST (baseline) on raw pixel features.

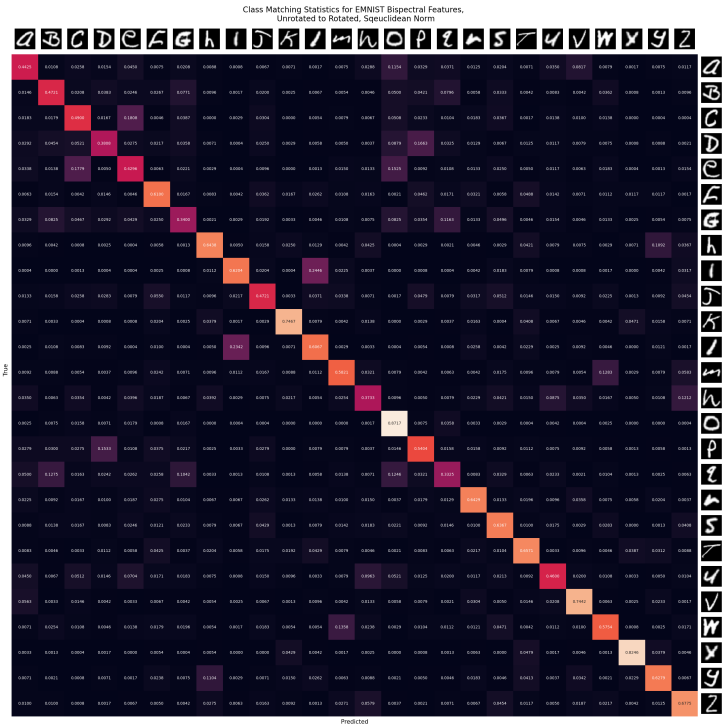


(a) L_1 Cost

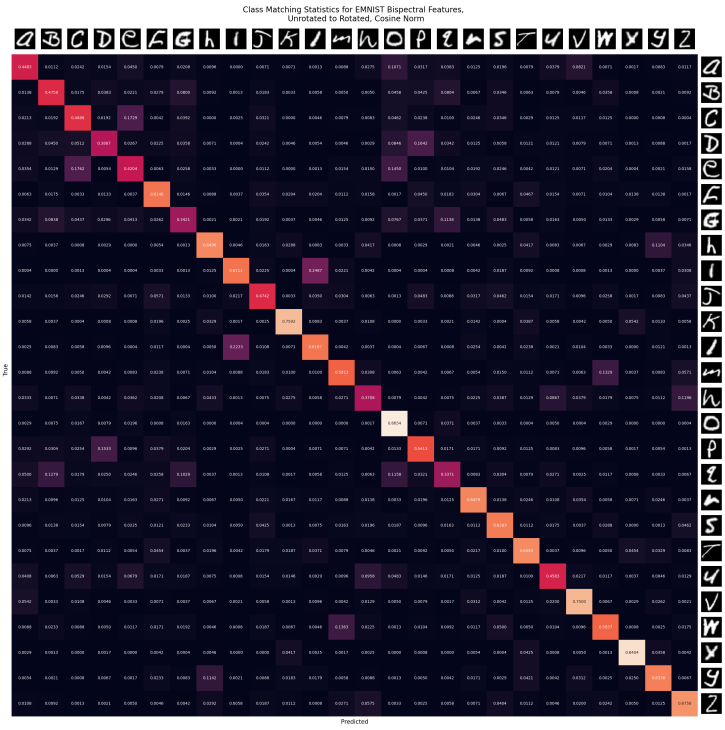


(b) L_2 Cost

Figure 19: Class matching statistics for OT plan from rotated EMNIST to unrotated EMNIST on bispectral features.

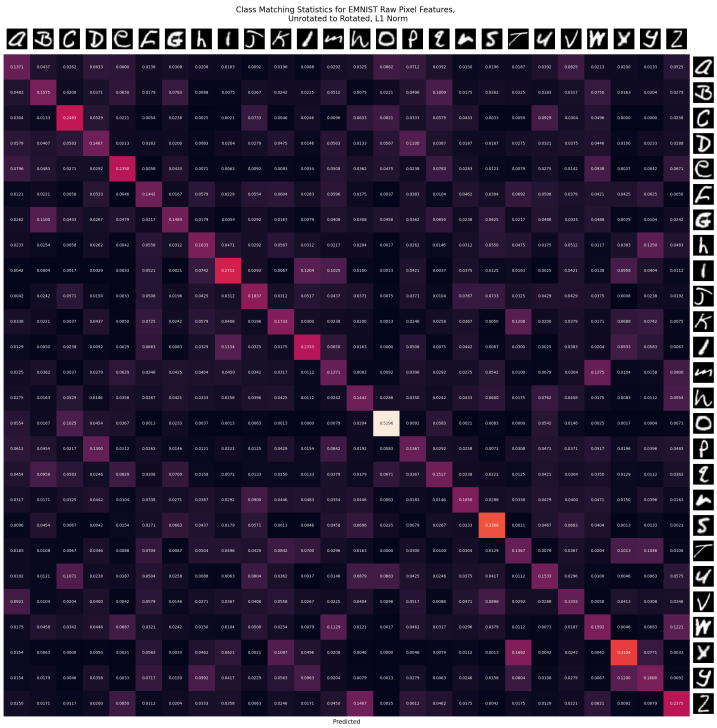


(c) L_2^2 Cost

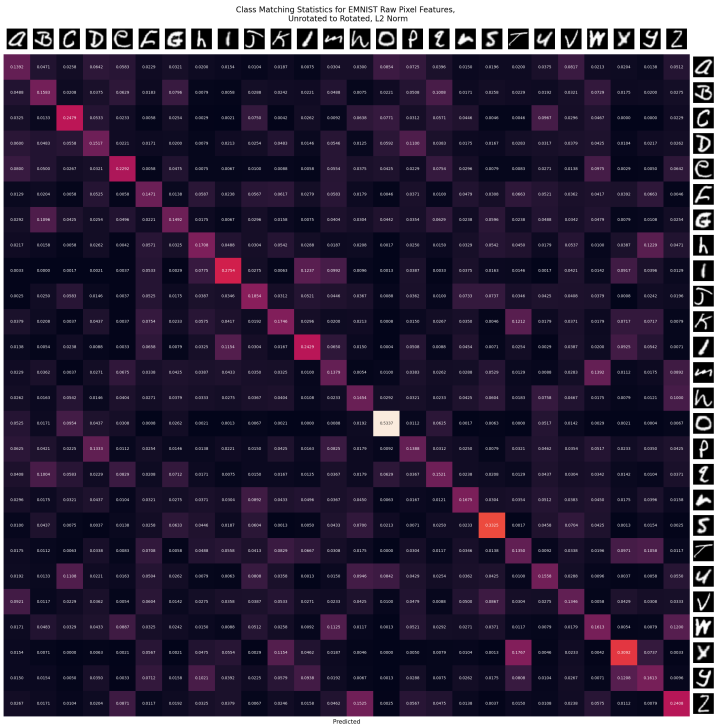


(d) cos Cost

Figure 19: Class matching statistics for OT plan from rotated EMNIST to unrotated EMNIST on bispectral features (continued).

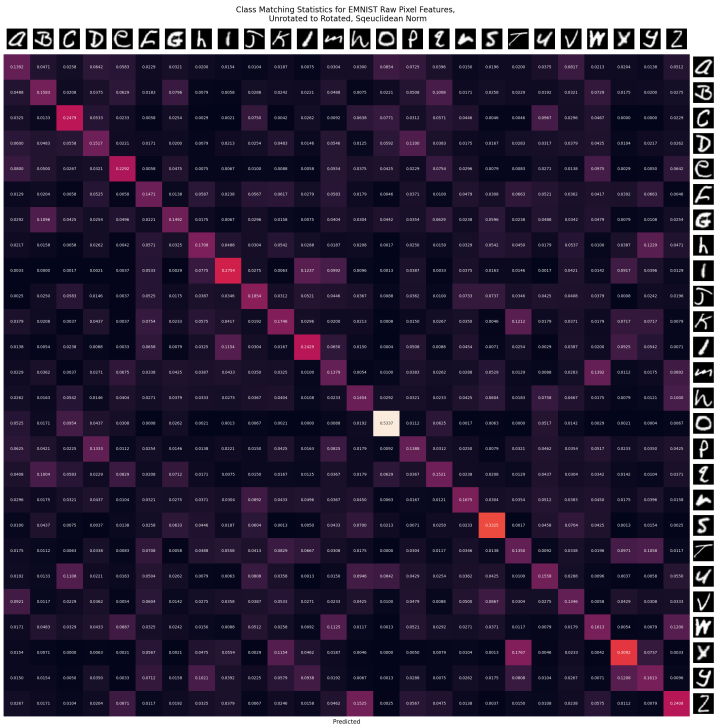


(*a*) L_1 Cost

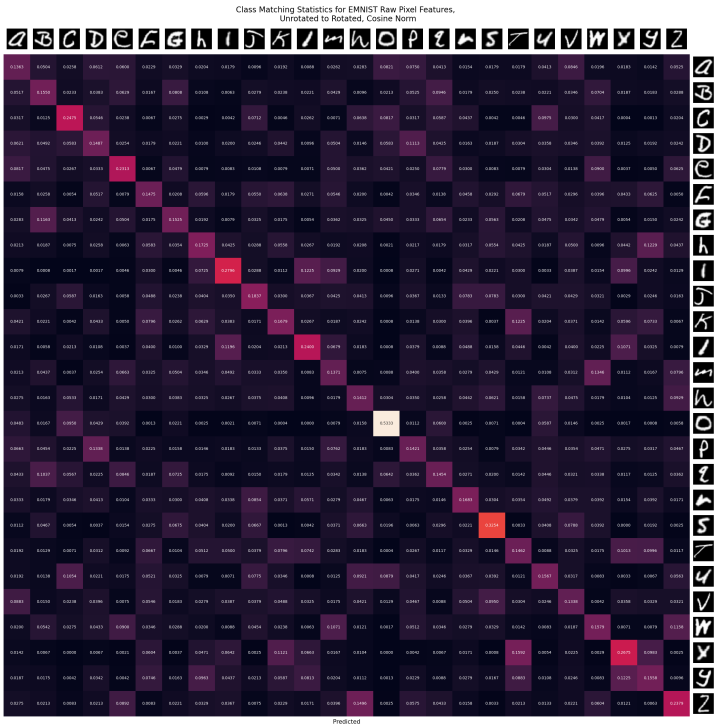


(*b*) L_2 Cost

Figure 20: Class matching statistics for OT plan from rotated EMNIST to unrotated EMNIST on raw pixel features.

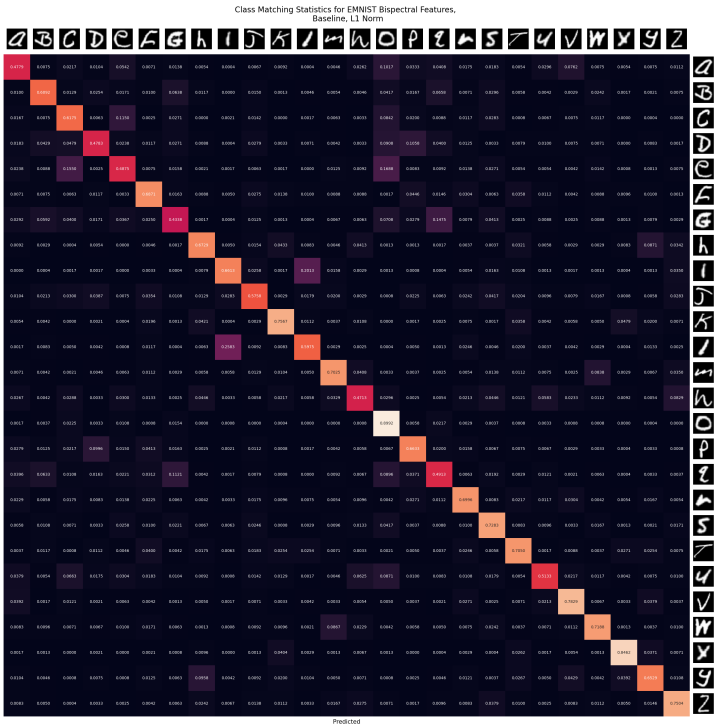


(c) L_2^2 Cost

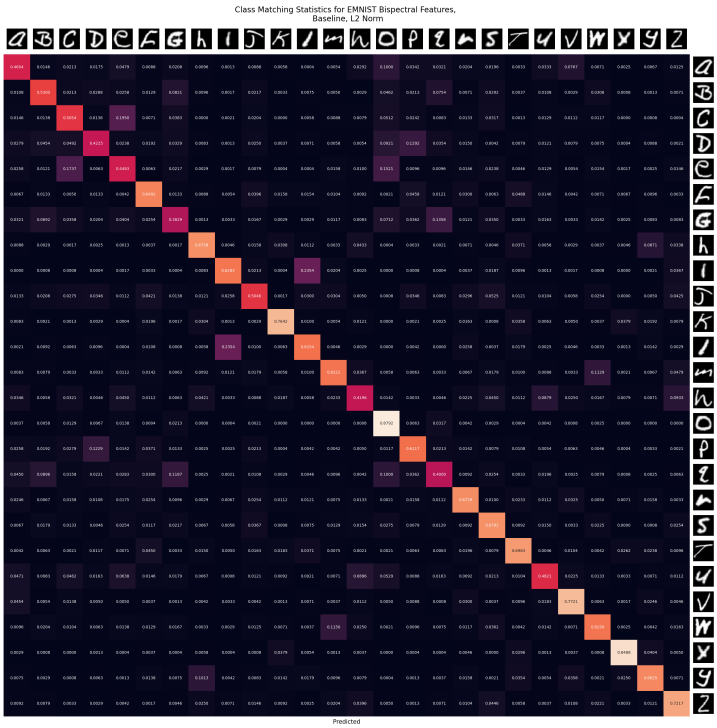


(d) cos Cost

Figure 20: Class matching statistics for OT plan from rotated EMNIST to unrotated EMNIST on raw pixel features (continued).

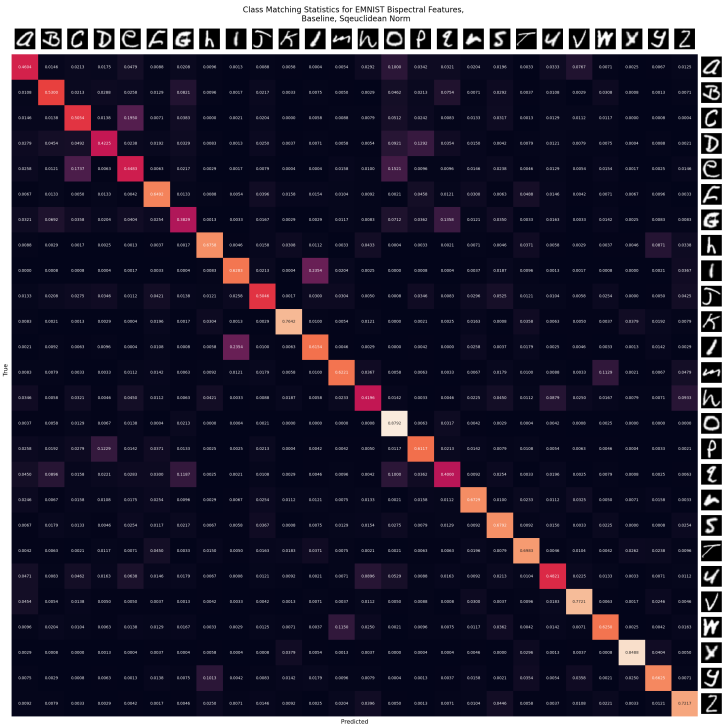


(*a*) L_1 Cost

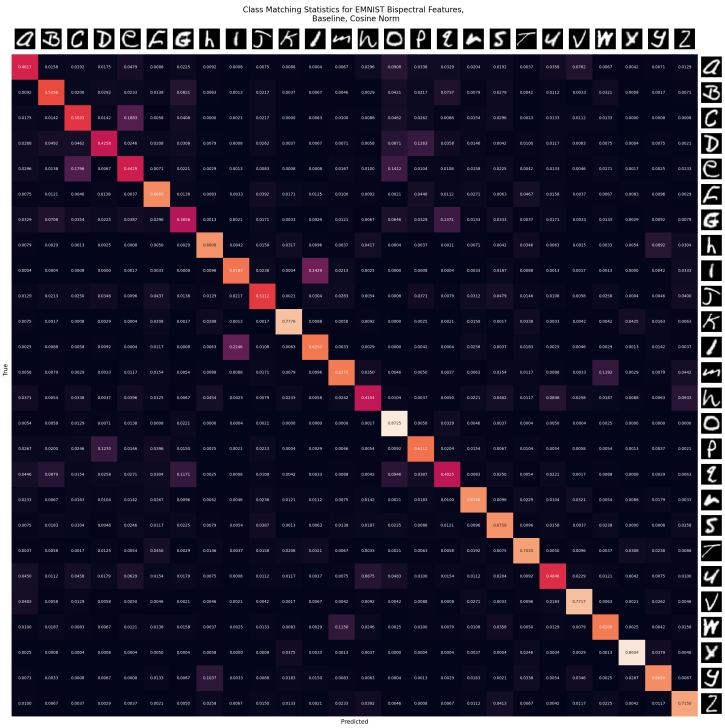


(*b*) L_2 Cost

Figure 21: Class matching statistics for OT plan from unrotated EMNIST to unrotated EMNIST (baseline) on bispectral features.

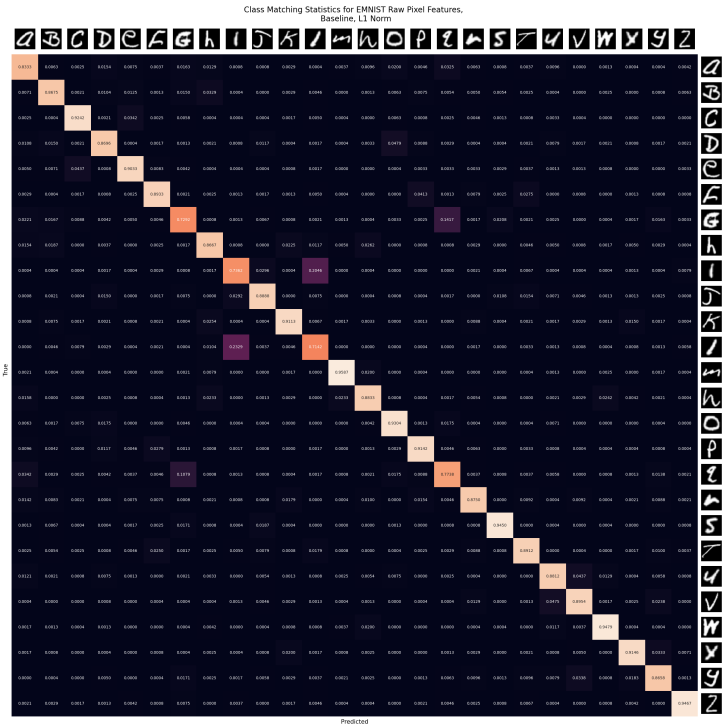


(c) L_2^2 Cost

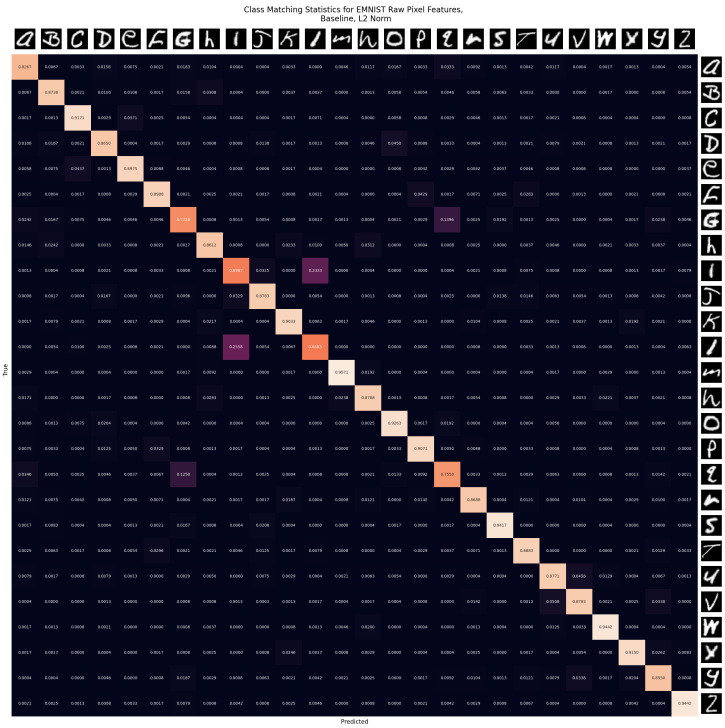


(d) cos Cost

Figure 21: Class matching statistics for OT plan from unrotated EMNIST to unrotated EMNIST (baseline) on bispectral features (continued).

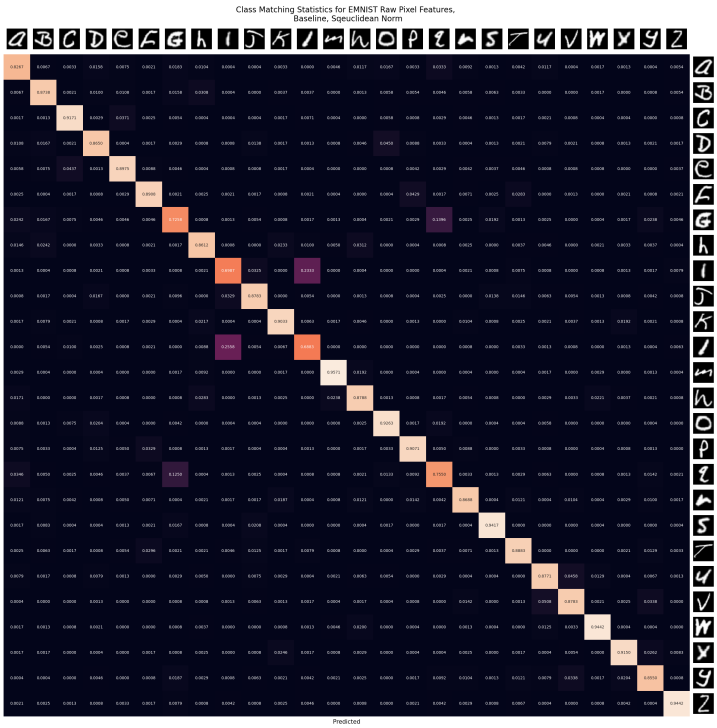


(*a*) L_1 Cost

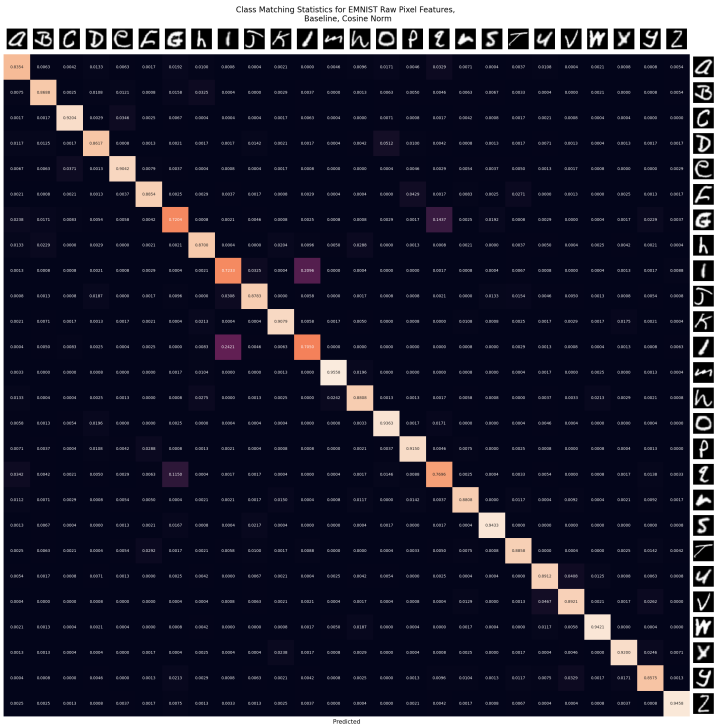


(*b*) L_2 Cost

Figure 22: Class matching statistics for OT plan from unrotated EMNIST to unrotated EMNIST (baseline) on raw pixel features.



(c) L_2^2 Cost



(d) cos Cost

Figure 22: Class matching statistics for OT plan from unrotated EMNIST to unrotated EMNIST (baseline) on raw pixel features (continued).