
TEN: TWIN EMBEDDING NETWORKS FOR THE JIGSAW PUZZLE PROBLEM WITH ERODED BOUNDARIES

Daniel Rika

Dept. of Computer Science
Bar-Ilan University
Israel, Ramat-Gan 52900
danielrika@gmail.com

Dror Sholomon

Dept. of Computer Science
Bar-Ilan University
Israel, Ramat-Gan 52900
dror.sholomon@gmail.com

Eli (Omid) David

Dept. of Computer Science
Bar-Ilan University
Israel, Ramat-Gan 52900
mail@elidavid.com

Nathan S. Netanyahu

Dept. of Computer Science
Bar-Ilan University
Israel, Ramat-Gan 52900
nathan@cs.biu.ac.il

ABSTRACT

This paper introduces the novel CNN-based encoder *Twin Embedding Network* (TEN), for the jigsaw puzzle problem (JPP), which represents a puzzle piece with respect to its boundary in a latent embedding space. Combining this latent representation with a simple distance measure, we demonstrate improved accuracy levels of our newly proposed pairwise *compatibility measure* (CM), compared to that of various classical methods, for degraded puzzles with eroded tile boundaries. We focus on this problem instance for our case study, as it serves as an appropriate testbed for real-world scenarios. Specifically, we demonstrated an improvement of up to 8.5% and 16.8% in reconstruction accuracy, for so-called Type-1 and Type-2 problem variants, respectively. Furthermore, we also demonstrated that TEN is faster by a few orders of magnitude, on average, than a typical deep neural network (NN) model, *i.e.*, it is as fast as the classical methods. In this regard, the paper makes a significant first attempt at bridging the gap between the relatively low accuracy (of classical methods) and the intensive computational complexity (of NN models), for practical, real-world puzzle-like problems.

1 Introduction

Numerous successful studies have been pursued over the years for solving the intriguing *jigsaw puzzle problem* (JPP). In particular, most advanced methods have focused recently on jigsaw puzzles containing non-overlapping square pieces. The reconstruction of such puzzles is composed typically of two main steps. First, a *compatibility measure* (CM) is computed for all pairwise piece sides, to obtain numerically the degree of compatibility for each possible piece pair. After computing all of the pairwise CMs, a reconstruction phase kicks in. The goal of any reconstruction algorithm is to determine correctly the location and orientation of each piece, relying solely on the CM scores computed in the first step. All previous works vary in the way they compute the pairwise CMs and in the way they try to assemble a puzzle as accurately as possible. Although the two steps are equally essential for the reconstruction, we focus here on the CM computation of all pairwise pieces, which could be highly expensive due to the order of $\mathcal{O}(N^2)$ operations required. (Note that an N -piece puzzle requires $16N^2$ pairwise CM computations.)

Classical, most commonly used CMs include the *sum of squared distances* (SSD) proposed by Cho *et al.* [1], the so-called *prediction-based compatibility* (PBC), employing $(L_p)^q$ variants due to Pomerantz *et al.* [2], the L_1 -norm compatibility of Paikin and Tal [3], and the *Mahalanobis gradient compatibility* (MGC) proposed by Gallagher [4].

These and other measures (*e.g.* [5, 6, 7]) yield high quality CMs, which allow for successful reconstruction of very large puzzles with up to tens of thousands of pieces.

The effectiveness of the above methods greatly depends, however, on the notion of a *synthetic puzzle*, *i.e.* a puzzle obtained by cropping a high-resolution image into N non-overlapping square pieces of size $P \times P$ pixels. (Such puzzles contain perfect pieces, and are considered information rich.) Indeed, the classical methods tend to be simple, fast, and very accurate under this favorable working assumption, which bodes well with CMs based mostly on piece boundary pixels.

When encountering, though, harder puzzle problems (due to various *real-world* scenarios, such as monochromatic puzzles, eroded piece boundaries, missing pieces, etc.), the classical measures tend to perform poorly, thereby affecting negatively the reconstruction outcome. Mondal *et al.* [8] were the first to handle this weak spot for puzzles containing *eroded boundaries*. They proposed a new pairwise CM called *weighted-MGC* (wMGC), which combines the SSD and MGC CMs. Despite their slight improvement, their wMGC still suffers from severe degradation in performance for eroded piece boundaries of 1–2 pixels.

The growing influence of *deep neural networks* (DNNs), in many related fields, has given rise recently also to the development of more accurate and robust CMs for harder variants of the JPP. Pai $\ddot{\text{x}}$ ao *et al.* [9] used a pre-trained SqueezeNet [10] model and fine-tuned it to obtain an effective CM for *strip-cut shredded documents*. Their NN-based pairwise CM enabled them to reassemble mechanically shredded documents with 94% accuracy. Another real-world JPP variant concerns the reconstruction of *Portuguese tile panels*, where often (square) real panel tiles have been severely degraded chromatically and eroded along their edges over hundreds of years. Rika *et al.* [11] proposed an ensemble of four NN-based models for dealing with this challenging variant. Their NN-based CM provides a significant improvement over classical methods, which made the reconstruction of Portuguese panels feasible for the first time. Regarding the problem of eroded boundaries, Bridger *et al.* [12] used a *generative adversarial network* (GAN) to “in-paint” the gap between pairs of pieces. After sufficient training of their generator, they keep its weights fixed and apply a second training stage to fine-tune the discriminator to differentiate between a true adjacent in-paint pair and false ones. Inspired by Pix2Pix [13], their generator uses a U-Net encoder-decoder architecture [14] and a Markovian discriminator with a 3-layer encoder. They carried out their work on a piece size 64×64 pixels, with an erosion of 2 or 4 boundary pixels, and reported superior results to those of the classical L_1 -norm compatibility. Other methods aimed at solving the JPP entirely by DNNs [15, 16] were shown to work only on virtually impractical puzzle sizes of 4×4 .

Unfortunately, DNN-based CMs consist of millions of parameters, such that each inference becomes very computationally-intensive. Since the model should be employed $16N^2$ times for an N -piece puzzle, the entire CM phase becomes highly infeasible in the lack of powerful dedicated hardware.

As shown in Section 3.4.3, the running time of a highly-intensive NN-based CM grows exponentially, and could take a couple of hours just to compute the pairwise CMs of a puzzle with only a few thousands pieces. Thus, tackling real-world problems like *shredded documents*, *ancient tile panels*, etc., which involve multiple puzzles containing millions of pieces, would render essentially NN-based CMs impractical.

To alleviate significantly this serious deficiency, we propose in this paper a new methodology for computing efficiently a pairwise CM via the *Twin Embedding (Neural) Network* (TEN) framework. The main idea of our novel approach is to employ TEN to represent an *entire puzzle piece* with respect to *each of its boundaries in a latent embedding space*. Combining this representation with a simple distance measure, we then demonstrate that our newly obtained CM is significantly more accurate than well-known classical CMs and, at the same time, is faster by a few orders of magnitude than typical NN-based CMs. The main highlights of our work are summarized as follows:

- Proposed a new NN-based framework and architecture to represent via embeddings an entire puzzle piece with respect to its boundaries;
- Demonstrated the use of embedding NNs for speeding up the CM computation by a few orders of magnitude, relatively to a standard end-to-end NN-based scheme;
- Presented extensive Top-1 accuracy (to be defined) and reconstruction results, which demonstrate TEN’s superior performance to that of most classical CMs;
- Conducted extensive empirical tests, involving 220 puzzles (with hundreds of pieces each) of three different datasets, as well as two reconstruction algorithms, to further establish the superior performance of our proposed framework, in comparison to previous CM methods.

We believe that the above contributions make a significant first step at bridging the gap between the relative low accuracy (of classical CMs) and the intensive computational complexity (of NN-based models), for various real-world puzzle-like problems.

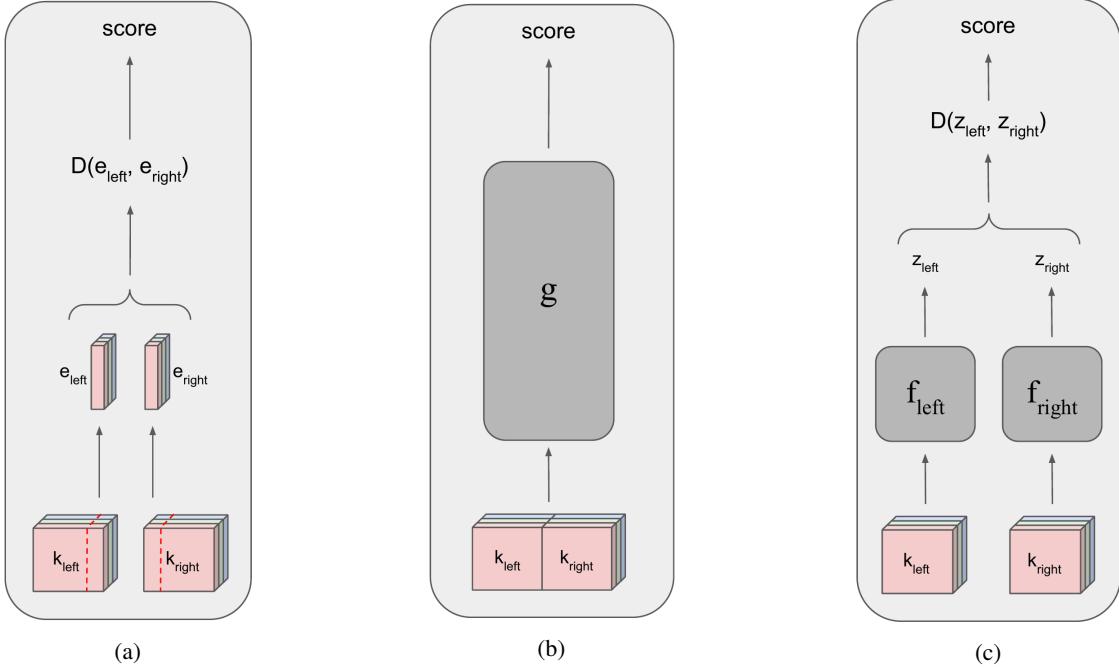


Figure 1: Schematic architectures for CM computation: (a) Traditional, (b) NN-based, and (c) TEN, where D represents a general distance measure function (not necessarily the same in (a) and (c)).

The rest of the paper is organized as follows. Section 2 describes in detail TEN’s framework and architecture. Section 3 presents our extensive experimental results, including a comparative performance evaluation with previous classical CMs. Finally, Section 4 makes concluding remarks and discusses future work.

2 TEN: Twin Embedding (Neural) Networks

Traditional methods (not based on NN models) typically compute the distance measure, $D(e_l, e_r) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, for computing the CM of a given pair of tiles, where e_l and e_r correspond, respectively, to the pixels along the left and right tile edges. Initially, only edge pixels were usually taken into account. More recent methods (like Gallagher’s MGC) considered 2-pixel wide edges for more informative processing. Hence, if a puzzle contains $P \times P \times C$ -size pieces, where P and C denote, respectively, the piece dimensions and number of (color) channels, then $e_l, e_r \in \mathbb{R}^{P \times 2 \times C}$. In contrast, NN-based dissimilarity functions work directly on an *entire tile pair* $[k_l : k_r]$, in an attempt to exploit maximum content information from the pieces k_l and k_r , which can be formulated as $g([k_l : k_r]) : \mathbb{R}^{P \times 2P \times C} \rightarrow \mathbb{R}$. Thus, the computational complexity associated with these functions is considerably more intensive.

Instead, we propose a hybrid approach, which uses an NN-based model for extracting the semantic information of a piece k into a latent space \mathcal{Z} , and combining it with a relatively simple distance measure D , to determine how two latent vectors are compatible to each other. The idea is that the NN-based model $f : \mathbb{R}^{P \times P \times C} \rightarrow \mathbb{R}^d$ extracts, in this manner, the most relevant features of the entire piece content with respect to each of its edges. Obviously, we would need two networks to obtain the above representations for a given pair of pieces, *i.e.* a left network f_{left} and a right network f_{right} for the left piece and right piece, respectively. Similarly to the traditional CM computations of a given pair, which work on edge pixels, we can now apply simply a distance measure $D(z_l, z_r) : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ with respect to the latent representations $z_l = f_{left}(k_l)$ and $z_r = f_{right}(k_r)$ of the pieces. (Note, that physically z_l corresponds to the right edge of k_l , and z_r corresponds to the left edge of k_r .) In our case, the dimensionality of the latent space is (expected to be) much smaller than that of the traditional approaches, as these methods operate on (all of) the raw edge pixels. See Figure 1 for the different schematic architectures of the above main approaches.

2.1 Architecture

As indicated before, TEN consists of twin CNN-based models, sharing the same architecture but with different learned weights. Each twin network of our small TEN version consists of four convolutional layers and two interleaving

max-pooling layers [17], using non-linear ReLU activation layers [18], and a fully-connected layer on top projecting the CNN outputs onto a latent vector in \mathcal{Z} . See Table 1 below for a detailed architecture.

Table 1: Our core architecture of both f_{left} and f_{right} embedding models; total number of model parameters depends on piece size and embedding dimension, *i.e.* (28, 28, 3) and $d = 40$, in our case; training our NN-based E2E model was done with input size of (28, 56, 3) and $d = 1$.

Layer Type	Output Dim	# Params
Input	(H, W, 3)	-
Conv2D, filter=(3, 3)	(H, W, 64)	1,728
Conv2D, filter=(3, 3)	(H, W, 128)	73,728
MaxPool2D, filter=(2, 2)	(H/2, W/2, 128)	-
Conv2D, filter=(3, 3)	(H/2, W/2, 256)	294,912
MaxPool2D, filter=(2, 2)	(H/4, W/4, 256)	-
Conv2D, filter=(3, 3)	(H/4, W/4, 512)	1,179,648
Linear	(d,)	32HWd
Total parameters:		1,550,016 + 32HWd

To extract even higher-quality embedding vectors, we also propose a larger version, TEN-Large, in which each twin is an ensemble of four CNN-based networks with the same architecture as in Table 1, but with different inputs. Following [11], each sub-network receives a unique color channel (*i.e.* red, green or blue), and the fourth network receives the entire (3-channel) RGB input. In principle, we observed that training an embedding model for each channel separately yields unique embedding features per channel, which results in added information to the ensemble architecture and in overall enhanced representation. TEN-Large has 20.4M parameters, *i.e.* 4 times as many as the the original TEN, as well as 4 times the number of embeddings. The pairwise compatibility using TEN-Large is given by

$$CM_{\text{TEN-Large}}(k_l, k_r) = \frac{1}{4} \sum_i D(f_{left}^i(k_l), f_{right}^i(k_r)), \quad (1)$$

where i represents a sub-network index of the ensemble.

2.2 Training with Embeddings

A procedure for computing a pairwise CM should return the lowest distance between an *anchor* boundary and its true adjacent piece boundary (*i.e.* *positive pair*), relatively to the distances between the *anchor* boundary and any piece edge in the puzzle (*i.e.* *negative pairs*). In order to train our twin embedding networks to do so, we used *triplet-loss* [19] defined by the following objective function:

$$\mathcal{L}_{\text{triplet}} = \max(0, D(z_a, z_p) - D(z_a, z_n) + \gamma), \quad (2)$$

where γ is set to 1, z_a denotes the embedding vector of f_{left} (as the anchor edge is always located on the left-hand side), and where z_p and z_n denote the embedding vectors of f_{right} (for the positive and negative edges, respectively), as they are located on the right-hand side. Plausible distance measures could be the inverse cosine similarity, L_1 distance, Euclidean distance, or any other distance measure between two vectors. In principle, we found that picking triplet piece edges from the same puzzle leads to better performance. This could be attributed to the notion that negative piece edges from the same puzzle are more likely to be similar to their positive counterparts, thereby forcing the twin embedding networks to extract more discriminative embedding vectors.

2.3 Post Processing

Note that TEN will return different dissimilarity scores for a given pair and its rotated version by 180° , since these pairs do not look the same from the network's perspective. Of course, obtaining an asymmetric compatibility measure for the very same pair does not make sense. To remedy this undesired effect, we apply post processing after computing all pairwise permutations (k_i, k_j) .

$$CM'(k_i, k_j) = \frac{CM(k_i, k_j) - \min(CM(k_i, *))}{\max(CM(k_i, *)) - \min(CM(k_i, *))} \quad (3)$$

We then ensure symmetry for the same pair of pieces by defining:

$$CM''(k_i, k_j) = CM''(k_j, k_i) = \frac{CM'(k_i, k_j) + CM'(k_j, k_i)}{2} \quad (4)$$

2.4 Type-1 and Type-2 Puzzles

It is common to consider two levels of difficulty for the JPP, namely *Type-1* and *Type-2*. Type-1 assumes that only piece locations are unknown, while their orientations are known and fixed. Type-2 relaxes this assumption, *i.e.* piece orientations, as well piece locations, are unknown. Note that a CM for Type-1 needs to discriminate, for each anchor piece k_a , between its positive piece k_p (*i.e.* ground-truth neighbor) and all of the negative candidates k_n , from a total number of $N - 1$ candidates. For Type-2, the total number of candidates rises to $4(N - 1)$, as there are 4 possible orientations for each piece. From a combinatorial point of view, there are essentially $4N$ anchor edges for an N -piece puzzle, such that each anchor edge has $N - 1$ and $4(N - 1)$ possible matching edges, for Type-1 and Type-2, respectively. In other words, a CM needs to be computed $4N(N - 1)$ and $16N(N - 1)$ times, *i.e.* roughly $4N^2$ and $16N^2$ times for Type-1 and Type-2, respectively.

2.5 Fast Inference

Let us now examine theoretically the significant reduction in time complexity due to the use of embeddings, as described in Section 2. Assume a puzzle of N pieces with unknown piece position and orientation. As previously indicated, there are $16N^2$ possible combinations for this Type-2 variant, *i.e.* a CM module must be executed for all of these $16N^2$ combinations. Computing CMs for this large number of combinations by the classical methods may not have been a major issue, due to their relatively simple computation. In contrast, NN-based CMs are very computationally intensive, which makes them rather slow and virtually impractical for puzzles containing more than a few hundred pieces. Fortunately, our proposed TEN framework offers a highly-efficient alternative. Observe that each of the f_{left} and f_{right} networks needs to execute only $4N$ times to capture all of the 4-edge representations of each puzzle piece.

After TEN is executed $4N$ times, we save all of these embedding representations in two tensors of size $(N, 4, d)$, T_{left} and T_{right} , where d is the embedding dimensionality. At this point, we merely employ a very simple distance measure on the embeddings to obtain the CMs of all pairwise puzzle pieces. Although the distance measure needs to be computed $16N^2$ times, the running time it takes is by far smaller than employing an NN-based CM the same number of times, which involves millions/billions of FLOPS (for each pair). In fact, the larger the puzzle is, the larger the speedup gain obtained over an NN-based CM, as will be demonstrated in Subsection 3.4.3.

3 Experimental Results

During our extensive experiments, we trained and evaluated the following three models: (1) Regular version of our TEN architecture, (2) ensemble version, TEN-Large, which contains four times as many parameters as TEN, and (3) pairwise, NN-based end-to-end (E2E) CM, as depicted in Figure 1(b). For the latter model, we used basically the same architecture presented schematically in Table 1, but instead of d outputs, this network outputs a single value, which indicates the degree of compatibility of a given pair of pieces.

We have experimented intentionally with an NN-based model, as well, to study not only the actual speedup gain due to TEN during inference, but to assess the expected accuracy gap as a performance reference for future research.

We trained all of the above networks with the same recipe, including an Adam optimizer [20], a learning rate of 1E-4 (with a decay factor of 0.9, in case the loss did not decrease for 5 epochs), a batch size of 64, and an epoch of 5000 iterations. To obtain the final results for TEN, TEN-Large, and NN-based E2E, we trained them over 600 epochs. TEN and TEN-Large were trained with *triplet-loss*, while the NN-based CM was trained with the *binary cross entropy loss* (labeling positive pairs as “0” and negative pairs as “1”).

3.1 Datasets

We used the 800 training images of the DIV2K [21] dataset to train our model. For the evaluation part, we used the 100 validation images of DIV2K, 100 images of the PIRM dataset [22], and 20 images of the MIT dataset (see Cho *et al.* [1]). We selected the above three different datasets, which are publicly available and commonly used in computer vision, to experiment with a decent, diverse collection of puzzles to establish our method’s reliability. To generate a degraded version of the original datasets, we then cropped each puzzle image into pieces of size 28×28 pixels, and deleted the boundary pixels of each piece, yielding an erosion effect of 13.8% (with respect to the piece size).

Table 2: Dimensionality effect: $d > 40$ does not seem to gain higher Top-1 accuracy for both distance measures used, running TEN on half epochs (rather than on full training phase).

Distance Measure	Embedding Size - d			
	10	20	40	80
1 – cosine	31.1%	37.9%	40.7%	40.5%
Euclidean distance	46.3%	51.4%	54.4%	54.7%

As indicated before, we experimented with eroded-boundary pieces, as this problem instance serves as a testbed for various practical, real-world JPPs, which must deal with damaged tiles around the edges. We generated only 1-pixel erosions, as this minimal degradation already resulted in poor CMs. (See Table 4 for specific results.) To evaluate the classical CMs on 1-pixel erosions, in a meaningful manner, we disposed the entire piece frame and considered only its remaining content (*i.e.*, instead of the original 28×28 -pixel pieces, we considered their corresponding eroded pieces of size 26×26 pixels). On the other hand, we retained the original piece size for all trainable CMs (*i.e.* TEN, TEN-Large, and NN-based E2E), such that each piece includes its eroded frame. Also, for the classical SSD, PBC, and L_1 CMs we retained the image representation in LAB color space (according to the specification of these methods), while using an RGB representation for all other CMs.

3.2 Dimensionality Effect

In general, an embedding dimensionality is one of its main characteristics. To determine practically the desired embedding dimensionality for an adequate representation of a piece edge e , we trained TEN on various dimensions, *i.e.* $d \in \{10, 20, 40, 80\}$. We stopped at $d = 80$, since TEN did not seem to yield better *Top-1 accuracy* on our test set, using a larger number of dimensions. (See exact definition of Top-1 accuracy in Subsection 3.4.1.) Indeed, empirical results consistently showed that using $d = 40$ was satisfactory, regardless of the distance measure used (see Table 2).

3.3 Distance Measure

Recall, computing the compatibility score between two embeddings z_l , z_r requires the use of a distance function $D : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$. Technically, D can be any desired function, including an NN-based one; however, to accelerate the computation as much as possible, we have focused on fast and simple distance measures.

- **Cosine similarity:** This embedding similarity is very common these days, in particular among the *self-supervised learning* field. Let z_l and z_r denote the embeddings obtained. Then the cosine similarity is given by

$$\text{cosine}(z_l, z_r) = \frac{z_l \cdot z_r}{\|z_l\| \|z_r\|} \quad (5)$$

In its raw form, the cosine projects all distances to the range $[-1, 1]$, where the closest embeddings are assigned a score of 1 and vice versa. To force the cosine similarity to assign a near-zero value to similar embeddings, we define

$$D(z_l, z_r) = 1 - \text{cosine}(z_l, z_r), \quad (6)$$

such that all distances now lie in the range $[0, 2]$.

- **L_p distance:** This is a family of distance measures which differ from one another by the parameter p . In this work we examined three p values $\in \{1, 2, 3\}$, which correspond to the well-known L_1 , L_2 (Euclidean distance), and L_3 metrics. In general, the L_p distance between two embeddings is calculated according to

$$L_p(z_l, z_r) = \left(\sum_{i=1}^d |z_l[i] - z_r[i]|^p \right)^{1/p} \quad (7)$$

In all of the experiments, we trained our models with the above distance measures for $d = 40$. A comparative performance is given in Table 3 with respect to Top-1 accuracy (see definition below). The results indicate that the L_p norm is preferable to the cosine similarity. More specifically, the best results were achieved using the L_2 Euclidean distance.

Table 3: Top-1 accuracies of our model (trained on half epochs) for four simple, fast distance measures and embedding size of 40; L_2 metric achieved highest score.

Distance Measure	Top-1 Accuracy
1 – cosine	40.7%
\underline{L}_1	51.8%
\underline{L}_2 (Euclidean distance)	54.4%
\underline{L}_3	54%

Table 4: Top-1 accuracies for various CMs: Our embedding-based CM is more accurate than all classical methods tested (best results appear in bold, and second best are underlined); for now, NN-based E2E yields higher rates.

Method	Type-1			Type-2		
	DIV2K	PIRM	MIT	DIV2K	PIRM	MIT
SSD	37.4%	41%	42%	29%	32.2%	33.5%
PBC	<u>46.4%</u>	<u>49.1%</u>	<u>43.7%</u>	<u>37.5%</u>	<u>39.3%</u>	<u>34.1%</u>
\underline{L}_1	47%	49.8%	46.6%	38.7%	40.7%	37.6%
MGC	<u>56.3%</u>	<u>59.5%</u>	<u>53.5%</u>	<u>45.4%</u>	<u>47.5%</u>	<u>41.1%</u>
TEN (ours)	<u>59.4%</u>	<u>61%</u>	<u>55.2%</u>	<u>50.5%</u>	<u>52.5%</u>	<u>46.6%</u>
TEN-L (ours)	64.6%	65.4%	59.1%	56.8%	57.5%	51.5%
NN-based E2E	71.4%	72.6%	69.4%	64.1%	64.2%	60.9%

3.4 Performance Results

3.4.1 Top-1 Accuracy

Ideally, given an anchor piece edge, we would like an algorithm that assigns the best pairwise CM to the true adjacent edge among all other piece edges in the puzzle. The most common criterion to quantitatively assess the performance of any such CM procedure is the Top-1 classification rate.

We define a pair (z_a, z_p) of an (embedded) anchor edge and its (embedded) true adjacent edge to be Top- i compatible (with respect to a given CM procedure) if its computed CM is greater than the i -th computed CM between the (embedded) anchor edge and any (embedded) negative edge $(z_a, z_n), \forall z_n \neq z_a, z_p$. Ideally, Top-1 should approach 100% accuracy for a perfect CM module. Practically, the higher it is, the more accurate the CM computation.

We computed Top-1 accuracies on the entire test set of 100 puzzles, each containing 800-1000 pieces. As the results indicate, NN-based E2E performs best, for now, with Top-1 accuracies of 71.4% and 64.1% for type-1 and type-2, respectively. Our TEN-Large yields, though, second-best CM scores with Top-1 accuracies of 64.6% and 56.8% for Type-1 and Type-2, respectively. This gap might be attributed to the intuitive insight that an NN-based model “sees” an entire pair before assessing the extent of compatibility between the two pieces. In contrast, the twin models f_{left} and f_{right} try to predict, essentially, each other’s embedding, which is a harder task.

3.4.2 Black-Box Reconstruction

Another way to conduct a comparative performance evaluation is to supplement each CM module with the same reconstruction algorithm, and test which CM, coupled with this reconstruction “black box”, contributes the most to the reassembled puzzle. We chose two reconstruction algorithms: (1) Gallagher’s greedy reconstruction algorithm, whose code is publicly available and which is commonly used in many other studies, and (2) the genetic algorithm (GA)-based reconstruction scheme due to [11]. Unlike Gallagher’s solver, the latter scheme searches for a global optimization by correcting piece misplacements at early stages of the reconstruction. To assess the reconstruction accuracy, we apply the *neighbor accuracy* criterion. (For the GA-based solver we actually report the average neighbor accuracy over 10 runs due to its non-deterministic behavior.) We experimented with all of the traditional CMs, as well as our proposed TEN and TEN-Large, and recorded for each method the average neighbor accuracy (over the test images of each dataset), after employing the above reconstruction solvers. Note that the reconstruction algorithms were employed uniformly, just for an additional comparative evaluation, *i.e.* the main point here is not necessarily the absolute neighbor

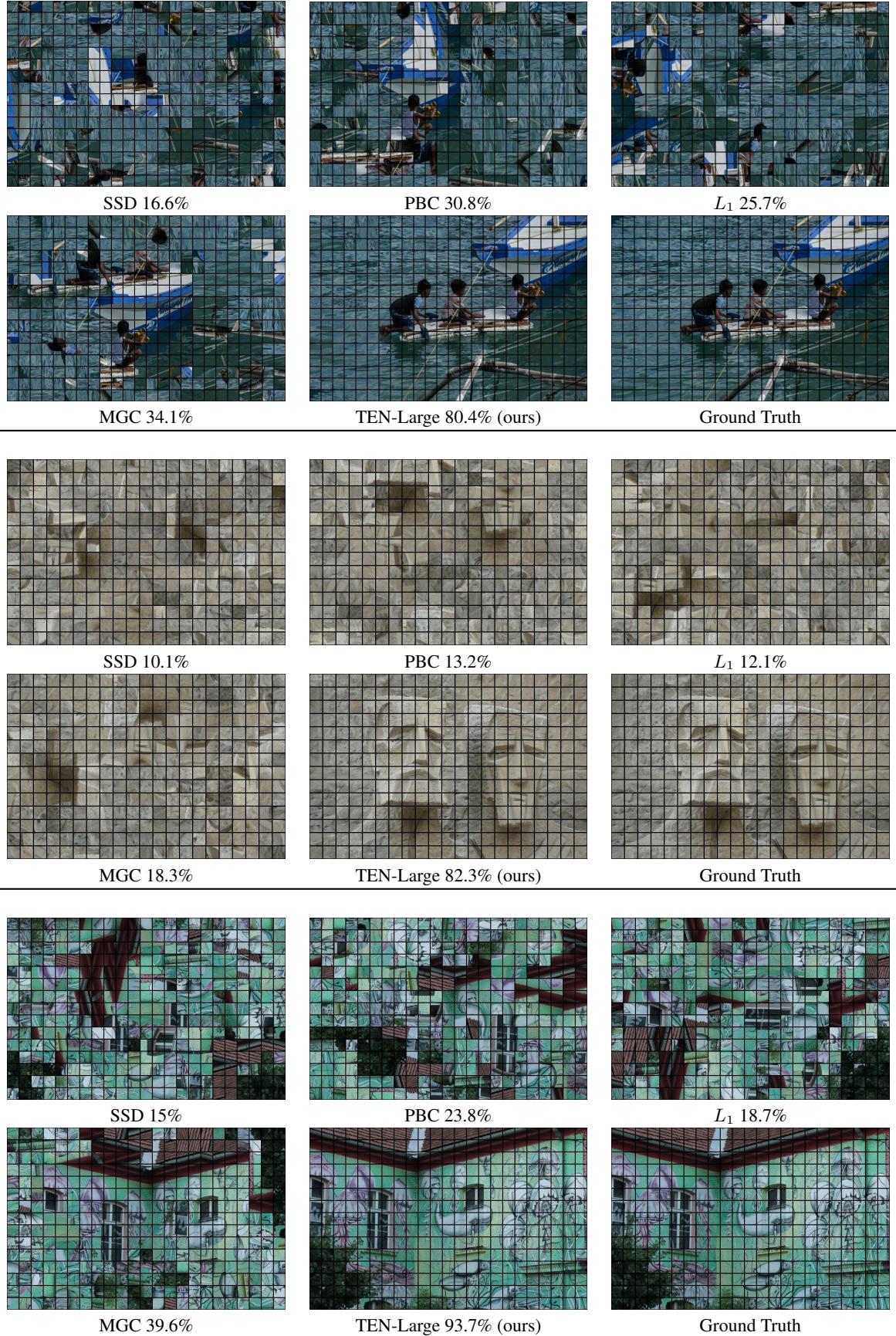


Figure 2: Sample of reconstructions due to [11] for Type-2 variant; our TEN-Large model exhibits superior performance to that of classical methods.

Table 5: Reconstruction neighbor accuracy due to evaluated CMs with Gallagher’s greedy solver [4] and GA-based solver [11]; average accuracy (over 10 runs for each puzzle) reported for latter reconstruction; TEN (and TEN-Large) models are superior to classical methods for both Type-1 and Type-2 (best results appear in bold, and second best are underlined).

Type-1						
Method	Gallagher’s solver			GA-based solver		
	DIV2K	PIRM	MIT	DIV2K	PIRM	MIT
SSD	28.7%	32.4%	34.5%	27.9%	34%	32.4%
PBC	43.7%	47.1%	36.4%	49.3%	53.2%	44.7%
L_1	42.2%	46%	41.3%	43.6%	50.9%	43.2%
MGC	<u>54.7%</u>	60.2%	<u>49.4%</u>	64.4%	71.4%	58.2%
TEN (ours)	52.9%	53.1%	44.9%	<u>67.5%</u>	<u>71.5%</u>	<u>61.2%</u>
TEN-Large (ours)	60.7%	<u>59.8%</u>	53.4%	72.9%	76.8%	65.1%
NN-based E2E	75.4%	76.1%	69%	82.5%	90.8%	87.3%

Type-2						
Method	Gallagher’s solver			GA-based solver		
	DIV2K	PIRM	MIT	DIV2K	PIRM	MIT
SSD	7.3%	9.8%	10.3%	16.3%	18.4%	18.3%
PBC	13.8%	16.5%	9.5%	26.1%	30.5%	20.9%
L_1	12.2%	15%	10.7%	25.6%	28.2%	22%
MGC	19.7%	22.7%	12.8%	40.2%	47.5%	32.3%
TEN (ours)	<u>21.8%</u>	<u>23.4%</u>	<u>16.6%</u>	<u>46.9%</u>	<u>50.9%</u>	<u>39.6%</u>
TEN-Large (ours)	30.2%	33.6%	22.7%	55.5%	59.4%	49.1%
NN-based E2E	48.5%	49.9%	39%	73.3%	79.7%	74.5%

Table 6: Wall-clock inference times (**in seconds**) of our proposed embedding-based CMs compared to NN-based E2E; TEN (and TEN-Large) boost inference speed by orders of magnitude as more pieces are added to puzzle.

# Pieces	TEN	TEN-Large	E2E
100	7.0×10^{-2}	1.8×10^{-1}	3.2×10^1
200	1.4×10^{-1}	3.9×10^{-1}	1.2×10^2
400	3.6×10^{-1}	8.9×10^{-1}	4.9×10^2
800	1.0×10^0	2.2×10^0	1.9×10^3
1600	3.2×10^0	6.2×10^0	7.9×10^3
3200	1.1×10^1	1.9×10^1	3.1×10^4

accuracy obtained, but rather the significant improvement due to the novel CM model introduced, relatively to the various traditional measures tested.

The neighbor accuracy results (for the evaluated CMs coupled with the reconstruction algorithms) are summarized in Table 5 for both Type-1 and Type-2 variants. In general, the results reveal the higher reconstruction accuracy due to the GA-based solver. More importantly, using TEN-Large leads to improved accuracy of up to 8.5% and 16.8%, respectively, for Type-1 and Type-2, relatively to the best known classical MGC.

A visual reconstruction comparison between the classical CMs and our best proposed model, TEN-Large, is shown in Figure 2 on a few selected images.

3.4.3 Inference Times

Table 6 gives the actual running times for computing all of the $16N^2$ combinations of a single puzzle with N pieces, on average. We experimented with $N \in \{100, 200, 400, 800, 1600, 3200\}$, i.e. with a growing number of pieces, to

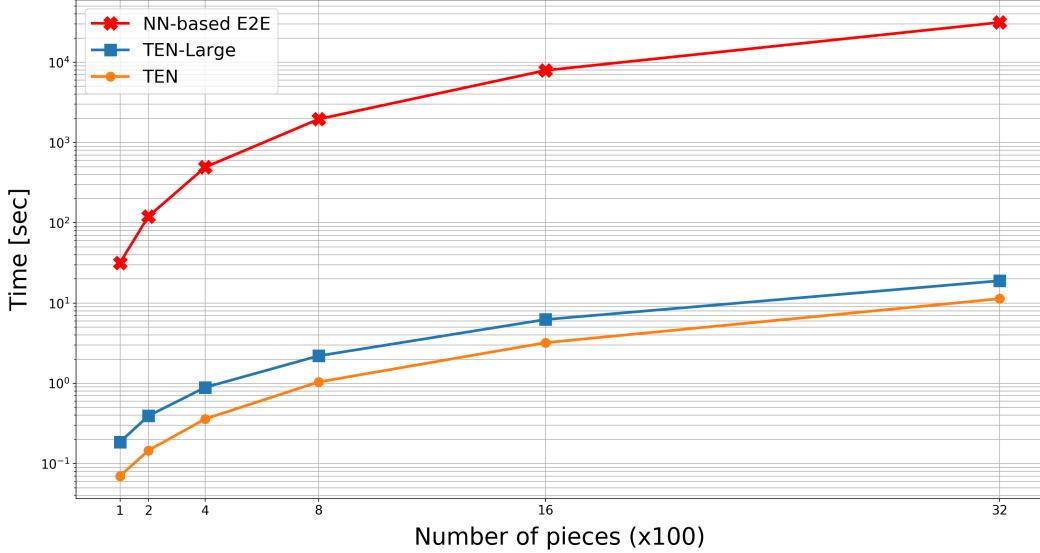


Figure 3: Comparative inference times of our new embedding-based CMs, TEN and TEN-Large, versus NN-based (E2E); computation time is dramatically reduced for $16N^2$ pairwise piece compatibilities.

underscore the impact of our proposed architecture compared to the NN-based (E2E) method. The running times due to the embedding (of TEN and TEN-Large) include the time to compute the final distance measures for all combinations.

Note the speedup gain as a function of puzzle size. Specifically, TEN/TEN-Large run 454/176 times faster than the NN-based model employed on 100-piece puzzles. Moreover, our TEN/TEN-Large models become 2771/1657 times faster than E2E on puzzles with 3200 pieces, which makes TEN the fastest known NN-based CM module. Figure 3 visualizes how fast our TEN and TEN-Large models compare to the NN-based (E2E) scheme.

Also, note that GPU/CPU utilization along with other factors like code optimization can cause changes in inference times as reported in Table 6. The reported inference times were measured on a modern PC with 3.5GHz CPU, 32GB RAM, and a single GPU with 11GB memory.

4 Conclusions

In this paper we presented a novel Twin Embedding (Neural) Network (TEN) framework, for computing very efficiently, yet fairly accurately, CMs for real-world JPP applications. The proposed embedding achieves a significant improvement in Top-1 accuracy compared to the classical CMs evaluated. This innovation yields also an enhanced reconstruction of puzzles with eroded-boundary tiles, as was demonstrated by the comparison between different CMs coupled with the same black-box reconstruction. We have focused deliberately on a problem domain with such degraded puzzles, unlike most JPP works, which deal typically with perfect, synthetic puzzles, since it offers a challenging testbed for real-world scenarios, where puzzle pieces undergo severe degradation along their edges.

We believe that our novel embedding paradigm could also affect significantly the accelerated research and development of further JPPs containing (tens of) thousands of eroded pieces for highly challenging problems, without compromising much on inference speed, as opposed to computationally-intensive NN-based CM models which run slower by a few orders of magnitude.

Although for now NN-based E2E models are relatively more accurate, we intend to address this limitation, as part of future work, to further close the accuracy gap while maintaining fast running times. For example, given the inherent redundancy between the f_{left} and f_{right} networks, in view of the same tasks performed by both models (albeit with respect to different edges), we intend to draw on this insight to derive a unified model for extracting bidirectional embeddings for a given piece.

References

- [1] T. S. Cho, S. Avidan, and W. T. Freeman. A probabilistic image jigsaw puzzle solver. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 183–190, 2010.
- [2] D. Pomeranz, M. Shemesh, and O. Ben-Shahar. A fully automated greedy square jigsaw puzzle solver. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9–16, 2011.
- [3] G. Paikin and A. Tal. Solving multiple square jigsaw puzzles with missing pieces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4832–4839, 2015.
- [4] A. C. Gallagher. Jigsaw puzzles with pieces of unknown orientation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 382–389, 2012.
- [5] K. Son, J. Hays, and D. B. Cooper. Solving square jigsaw puzzles with loop constraints. In *Proceedings of the European Conference on Computer Vision*, pages 32–46. Springer, 2014.
- [6] R. Yu, C. Russell, and L. Agapito. Solving jigsaw puzzles with linear programming. In *Proceedings of The British Machine Vision Conference*, pages 139.1–139.12, 2016.
- [7] S. Kilho, D. Moreno, J. Hays, and D. B. Cooper. Solving square jigsaw puzzle by hierarchical loop constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2222–2235, 2019.
- [8] D. Mondal, Y. Wang, and S. Durocher. Robust solvers for square jigsaw puzzles. In *Proceedings of the International Conference on Computer and Robot Vision*, pages 249–256, 2013.
- [9] T. M. Paixao, R. F. Berriel, M. C. Boeres, C. Badue, A. F. De Souza, and T. Oliveira-Santos. A deep learning-based compatibility score for reconstruction of strip-shredded text documents. In *Proceedings of the IEEE 31st SIBGRAPI Conference on Graphics, Patterns and Images*, pages 87–94, 2018.
- [10] N. I. Forrest, H. Song, W.M. Matthew, A. Khalid, J.D. William, and K. Kurt. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv:1602.07360*, 2015.
- [11] D. Rika, D. Sholomon, O. E. David, and N. S. Netanyahu. A novel hybrid scheme using genetic algorithms and deep learning for the reconstruction of Portuguese tile panels. In *Proceedings of the ACM Conference on Genetic and Evolutionary Computation*, pages 1319–1327, 2019.
- [12] D. Bridger, D. Danon, and A. Tal. Solving jigsaw puzzles with eroded boundaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3526–3535, 2020.
- [13] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [14] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. *arXiv:1505.04597*, 2016.
- [15] Marie-Morgane Paumard, David Picard, and Hedi Tabia. Deepzzle: Solving visual jigsaw puzzles with deep learning and shortest path optimization. *IEEE Transactions on Image Processing*, 29:3569–3581, 2020.
- [16] Ru Li, Shuaicheng Liu, Guangfu Wang, Guanghui Liu, and Bing Zeng. JigsawGAN: Auxiliary learning for solving jigsaw puzzles with generative adversarial networks. *IEEE Transactions on Image Processing*, 31:513–524, 2022.
- [17] J. Nagi, F. Ducatelle, G.A. Di Caro, D. Cireşan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L.M. Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *Proceedings of the IEEE International Conference on Signal and Image Processing Applications*, pages 342–347, 2011.
- [18] V. Nair and G.E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, 2010.
- [19] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.
- [20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2017.
- [21] E. Agustsson and R. Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, July 2017.
- [22] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor. 2018 PIRM challenge on perceptual image super-resolution. *CoRR*, 2018.