Structure-aware Data Consolidation (Supplementary Material)

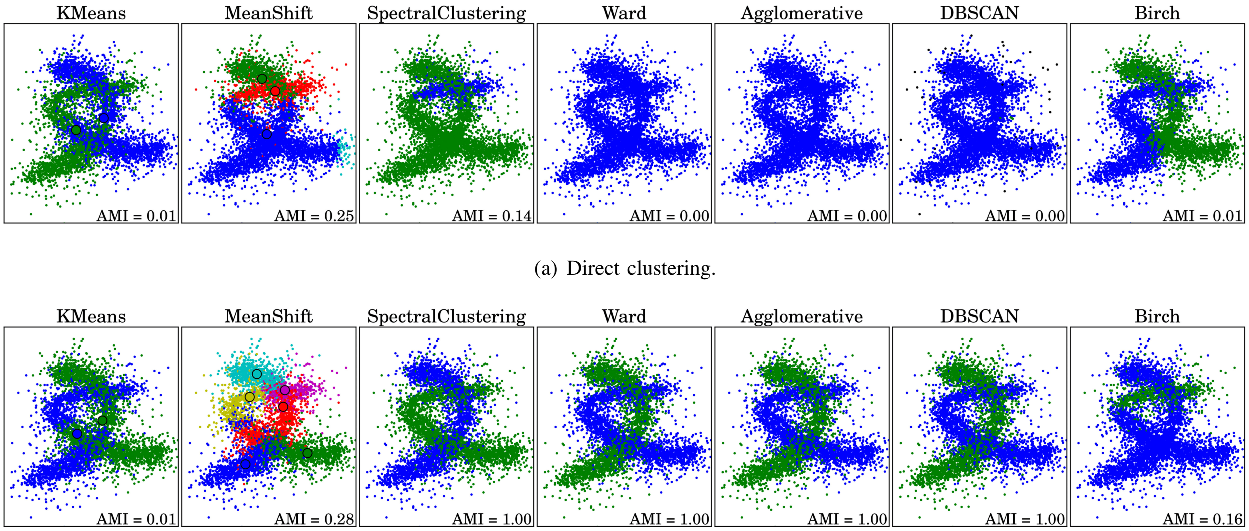
Shihao Wu, Peter Bertholet, Hui Huang, *Member, IEEE*
Daniel Cohen-Or, Minglun Gong, *Member, IEEE,* Matthias Zwicker, *Member, IEEE*

This supplementary material provides more experimental results for our structure-aware filtering (SAF) technique, and comparisons to state of the art manifold denoising methods (manifold denoising, MD [1], and manifold frequency denoising, MFD [2]). Figure 1 gives a wider range of comparison as a supplement of Figure 1 in the paper. Figure 2 justifies our choice of using median instead of mean repulsion. Figure 3 demonstrates that our theoretical analysis can guide the selection of parameter μ for noisy data in different dimensions, as a supplement of Figure 6 in the paper. Figure 4 illustrates a parameter selection strategy guided by our convergence criterion. Figure 5 further discusses parameter selection using the 3D data set from Figure 1 in the paper. Figure 6 illustrates how spectral clustering benefits from our data consolidation technique. Figure 7 illustrates the performance of our consolidation with increasing data density. Figure 8 shows additional clustering examples of low dimensional structures corrupted with high dimensional noise. Figure 9 shows additional examples of dimensionality reduction, including comparisons with MD and MFD. Figure 10 illustrates SAF performance with increasing dimensionality of the data, as a supplement of Figure 9 in the paper. Figures 11, 12 and 13 illustrate SAF performance with increasing noise level, as a supplement of Figure 10 in the paper. Figures 14 and 15 are supplements of Figure 11 (MINST) and Figure 12 (Yale Face) in the paper, respectively. These figures all also include comparisons with MD and MFD.

References

[1] M. Hein and M. Maier, "Manifold denoising," in *Advances in Neural Information Processing Systems*, 2007, pp. 561–568.

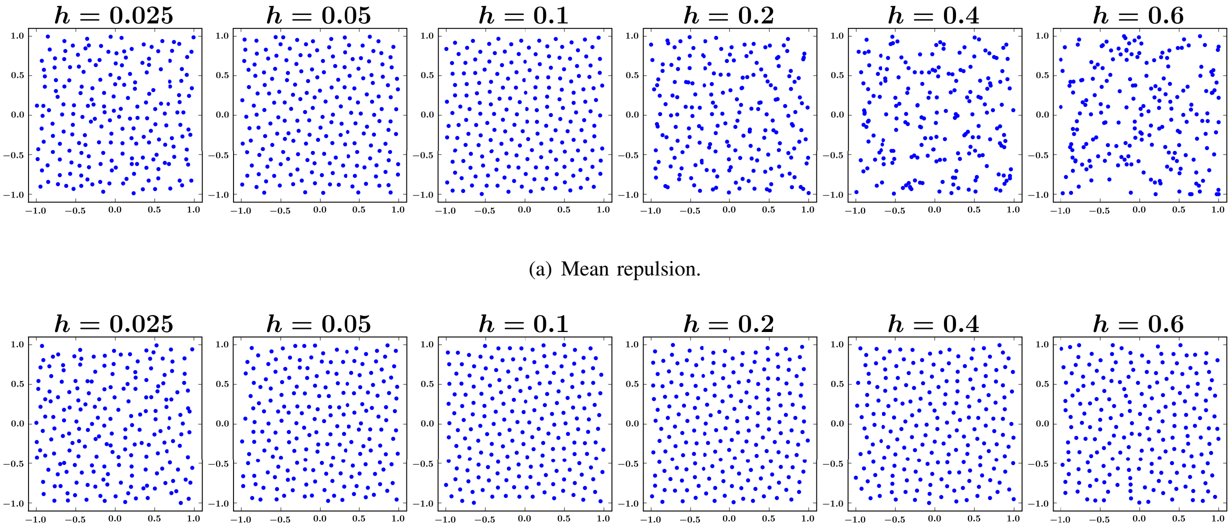
[2] S. Deutsch, A. Ortega, and G. Medioni, "Manifold denoising based on spectral graph wavelets," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2016, pp. 4673–4677.



(a) Direct clustering.

(b) Clustering with SAF.

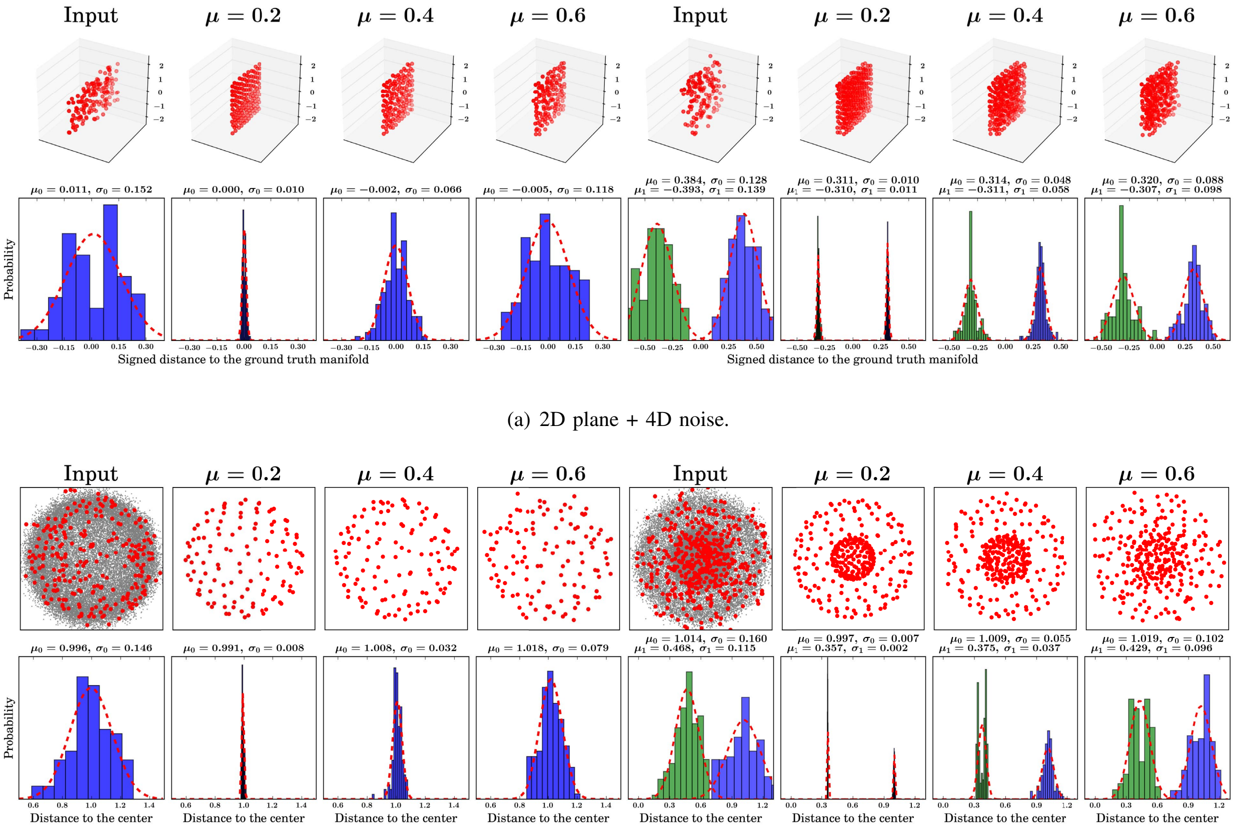
Fig. 1. Improved performance of additional clustering techniques on the data from Figure 1 in the paper. Directly using standard clustering techniques (a) may generate incorrect results when the clusters are intertwined and corrupted with noise. By projecting data points to the lower dimensional curve structures and filtering out the noise, we significantly improve clustering performance (b). Some techniques, such as k-means and mean shift clustering, do not profit from our approach, however, because they inherently cannot detect clusters with such elongated, curvy shapes. We also provide the adjusted mutual information (AMI) of the clustering results.



(a) Mean repulsion.

(b) Median repulsion.

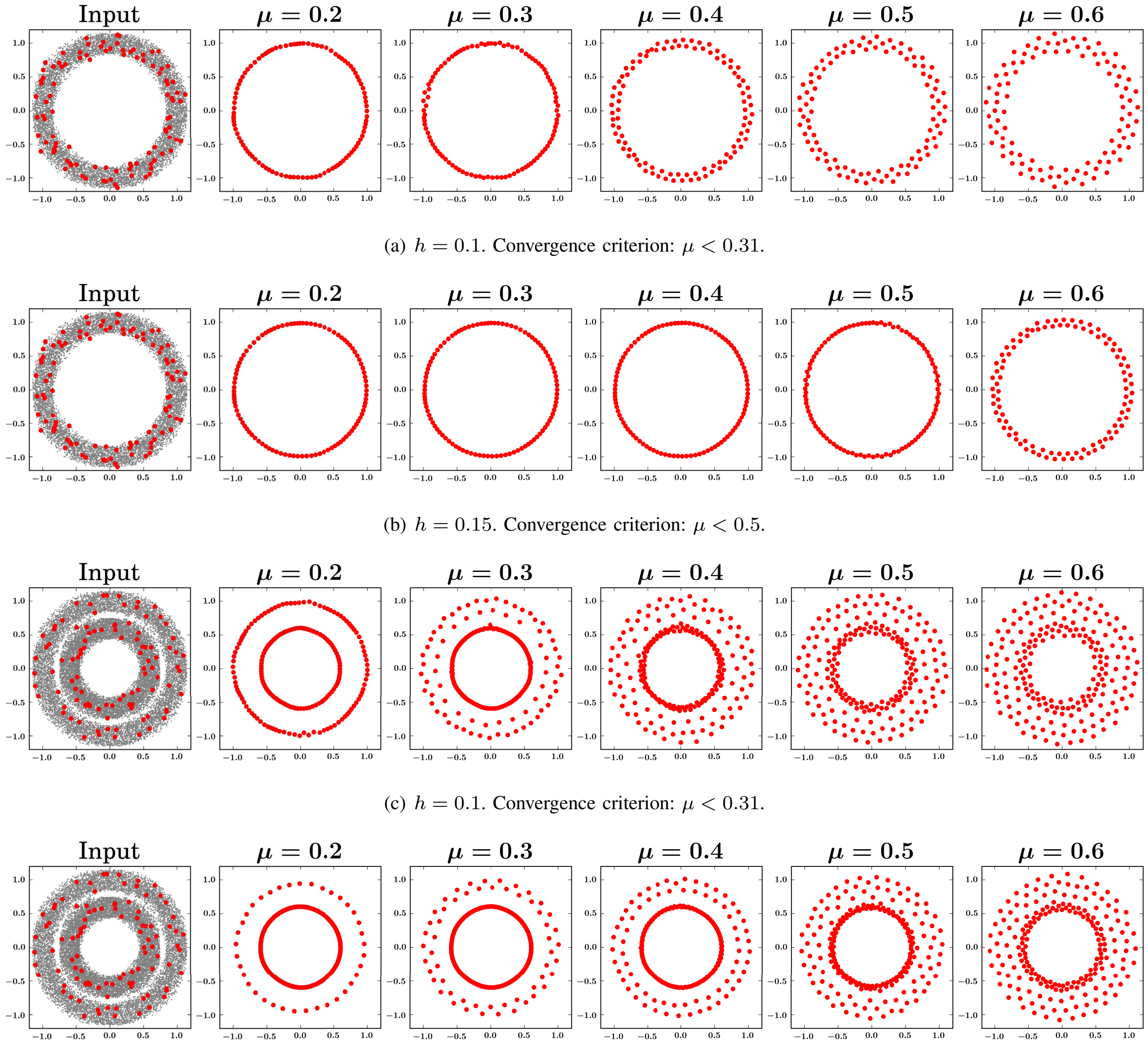
Fig. 2. We compare mean (a) and median (b) repulsion. We regularize a set of 2D points using only repulsion without data term, and with periodic boundary conditions. We compare the results with different h values. While both mean and median repulsion have similar convergence behavior, median repulsion is more robust and less sensitive to the neighborhood size h, which helps to generate locally uniform distributions.



(a) 2D plane + 4D noise.

(b) 3D sphere + 3D noise.

Fig. 3. Our theory can predict the convergence for high dimensional data. We add Gaussian noise with $\sigma = 0.15$ to 2D planes (a) and 3D spheres (b), set the kernel size $h = 0.1$, and show the results with different μ values applied. Our theoretical analysis predicts the convergence when $\mu < 0.31$. In each experiment, we test two types of input, i.e., single and double manifolds. At the bottom rows, we also demonstrate histograms of distances to the center of 2D planes and 3D spheres.



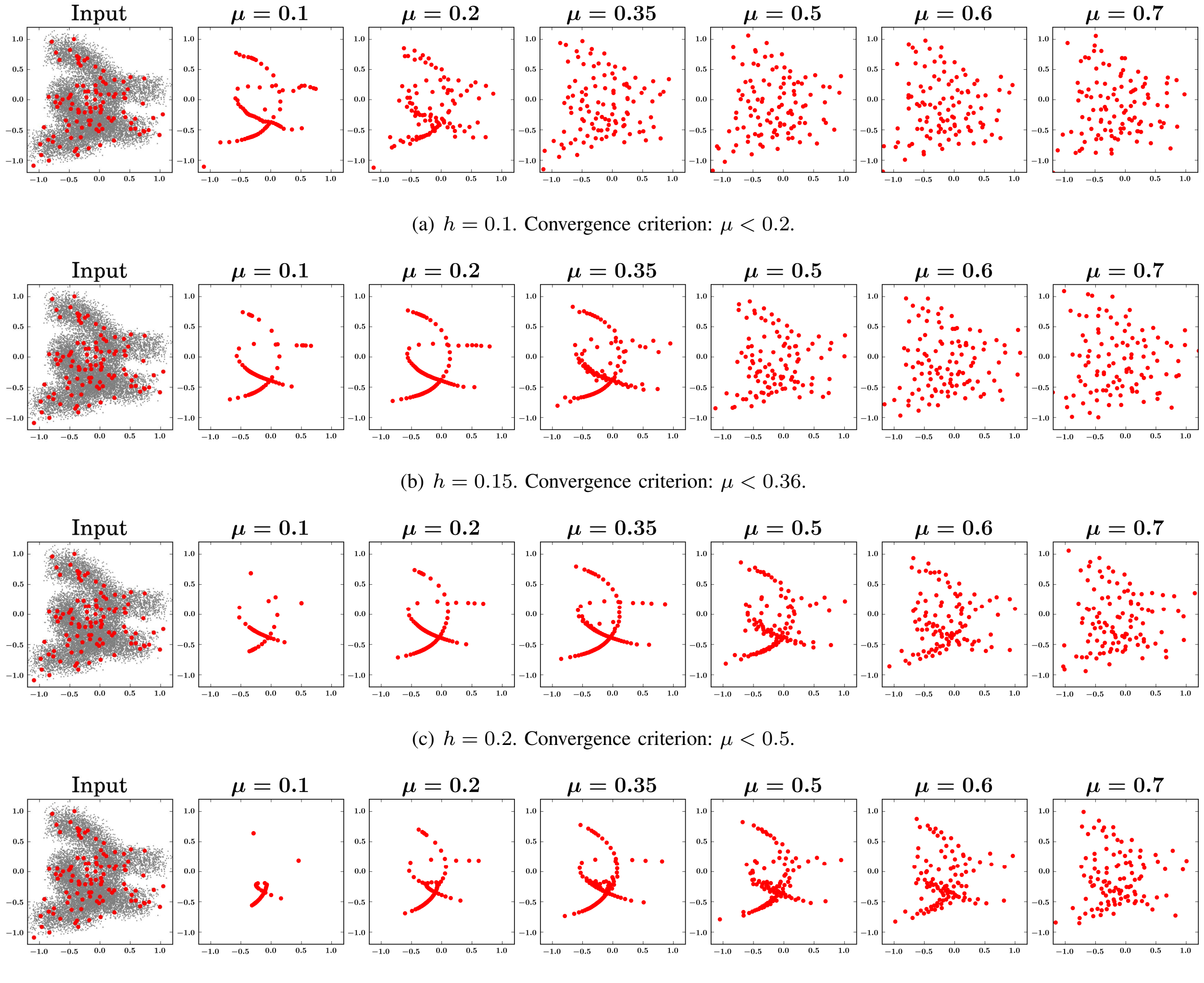
(a) $h = 0.1$. Convergence criterion: $\mu < 0.31$.

(b) $h = 0.15$. Convergence criterion: $\mu < 0.5$.

(c) $h = 0.1$. Convergence criterion: $\mu < 0.31$.

(d) $h = 0.15$. Convergence criterion: $\mu < 0.5$.

Fig. 4. We illustrate a parameter selection strategy guided by our convergence criterion. All input data (gray) are corrupted by Gaussian noise with $\sigma = 0.15$. For simple structures, such as the ring data in the first two rows, using a larger h value (second row) provides a smoother approximation of the underlying data density (gray). This encourages a more uniform distribution of output points and allows for a wider choice of repulsion strengths μ. For data with close-by structures, such as the two rings used in the last two rows, using a too large h value (second row) may be incapable of separating the nearby structures, however. In practice, one should first select an appropriate kernel size h according to the input data, and then choose μ close to the theoretical maximum to get a regular output point distribution. Note that we used denser input points (in gray color) than the consolidated output points (in red color) to better approximate our continuous formulation.



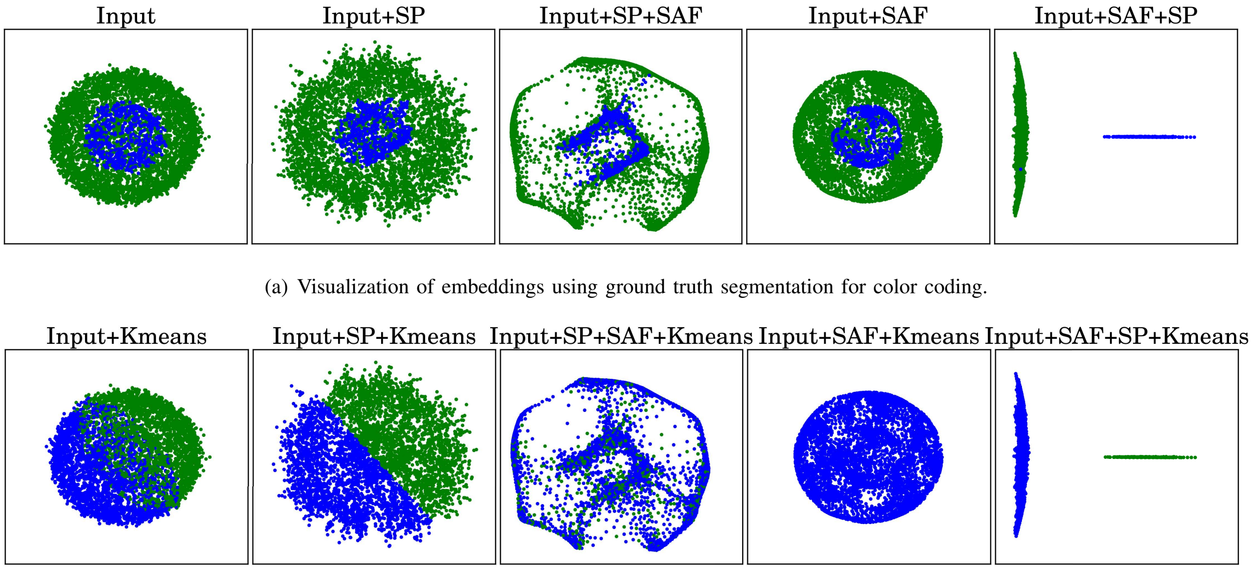
(a) $h = 0.1$. Convergence criterion: $\mu < 0.2$.

(b) $h = 0.15$. Convergence criterion: $\mu < 0.36$.

(c) $h = 0.2$. Convergence criterion: $\mu < 0.5$.

(d) $h = 0.25$. Convergence criterion: $\mu < 0.6$.

Fig. 5. We further discuss parameter selection using the 3D data set from Figure 1 in the paper. All input data are corrupted by Gaussian noise with $\sigma = 0.2$ shown in gray. Different h values are used for different rows. Using a big h value can better deal with noise and outliers, but different structures could be mistakenly merged if a too big h value is used (last row). On the other hand, using a small h value can better preserve the shape of the structure, but is less robust to the noise and may lead to less than satisfactory output distributions (red). We further show different μ values in each row. Using a big μ value usually gives more regular output distribution, but may violate the convergence criterion. Using a small μ value may cause disconnection in the structure. In general, guided by our theoretical analysis and knowledge of the data (e.g. how close nearby structures could be), as large as possible h and μ values are advised. We used denser input points (gray) than consolidated output points (red) to better approximate the continuous setting.



(a) Visualization of embeddings using ground truth segmentation for color coding.

(b) Visualization of different clustering strategies.

Fig. 6. We illustrate how spectral clustering, implemented via spectral embedding followed by k-means clustering, benefit from our data consolidation technique. The input data consists of two noisy, concentric 3D spheres with different radii corrupted with 6D noise. The top left shows an orthogonal projection of the input data to 2D. We compare five strategies for preprocessing and embedding the data before clustering (from left to right; the color codings in the top row show the ground truth clusters, the second row shows actual clustering results). All strategies rely on k-means clustering as their last step, and the results are shown in the bottom row. The first column, "Input" applies k-means directly to the input data; "Input+SP" is traditional spectral clustering, i.e., embedding using two eigenvectors of the affinity matrix followed by k-means; "Input+SP+SAF" uses the same spectral embedding, but consolidates before k-means clustering; "Input+SAF" consolidates and applies k-means both in 6D; "Input+SAF+SP" consolidates in 6D, and then uses a spectral embedding, following with k-means clustering. The experiment shows that the spectral embedding is sensitive to noise, and without our consolidation (second column from left) it fails to separate the clusters. In contrast, consolidation followed by spectral embedding (rightmost column) clearly separates the clusters and leads to correct results. Clustering in 6D without spectral embedding (first and fourth column from left) fails because k-means cannot separate the spherical shells in 6D.

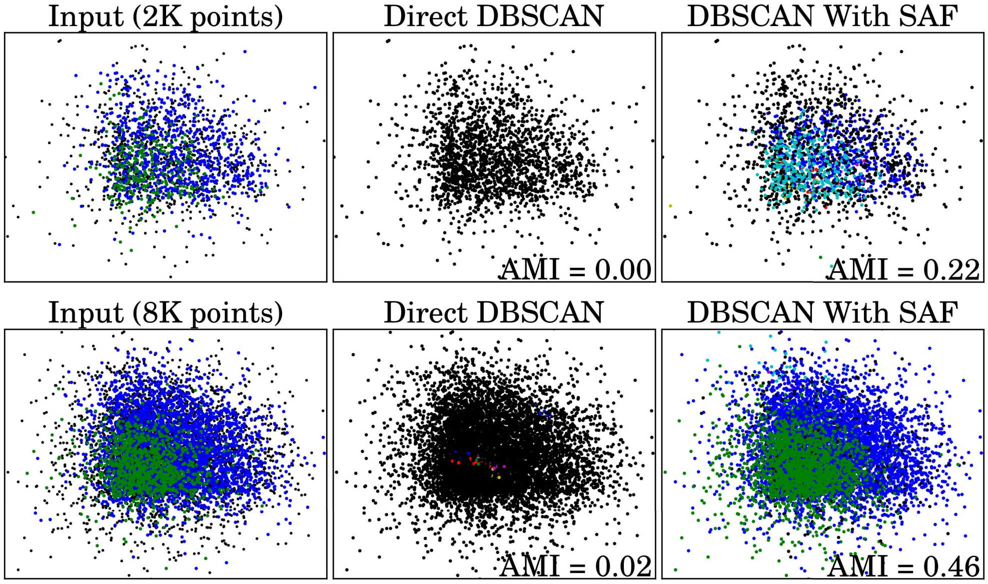
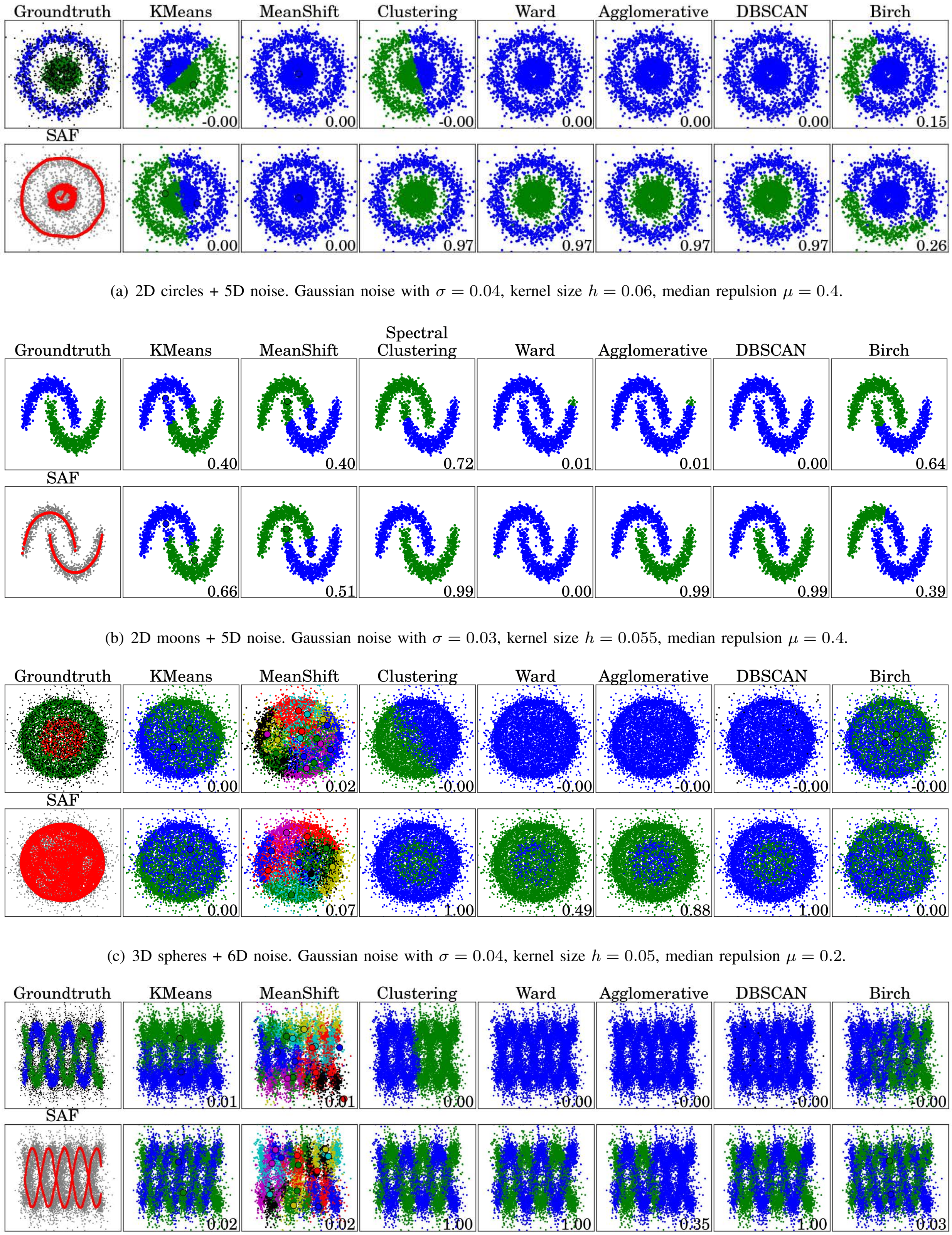


Fig. 7. We illustrate the performance of our consolidation with data density increasing. The data consists of two patches of two 6D concentric hyperspheres with different radii and corrupted with Gaussian noise. The two rows show clustering on sparser and denser inputs, respectively. As the dimensionality increases, the denser input is required to realize the effectiveness of consolidation. We use Gaussian noise $\sigma = 0.05$, kernel size $h = 0.024$ (relative to bounding box diagonal), $\mu = 0.25$, anisotropic SAF, and cluster in 6D without dimensionality reduction.



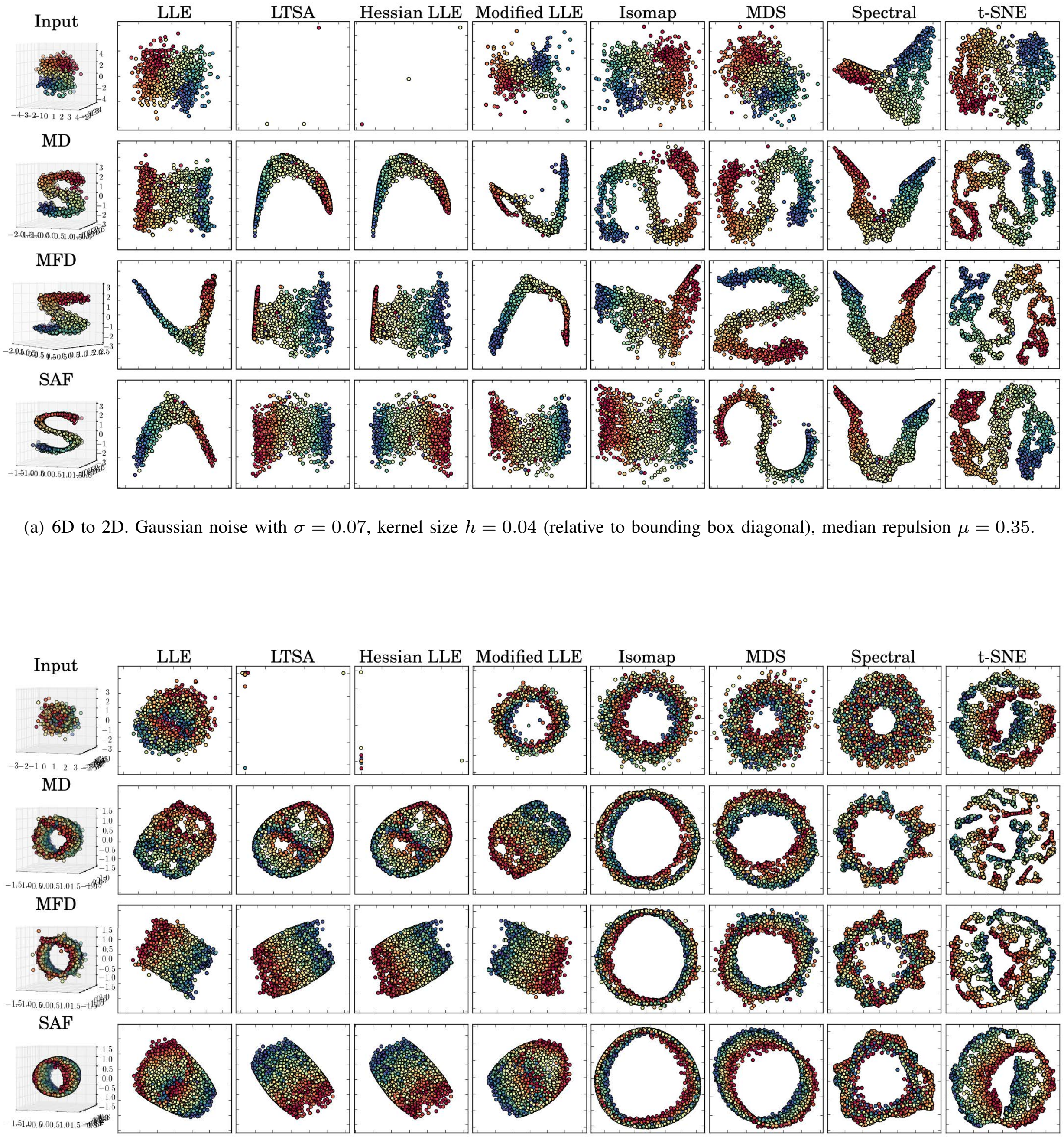
(a) 2D circles + 5D noise. Gaussian noise with $\sigma = 0.04$, kernel size $h = 0.06$, median repulsion $\mu = 0.4$.

(b) 2D moons + 5D noise. Gaussian noise with $\sigma = 0.03$, kernel size $h = 0.055$, median repulsion $\mu = 0.4$.

(c) 3D spheres + 6D noise. Gaussian noise with $\sigma = 0.04$, kernel size $h = 0.05$, median repulsion $\mu = 0.2$.

(d) 3D spirals + 10D noise. Gaussian noise with $\sigma = 0.025$, kernel size $h = 0.04$, median repulsion $\mu = 0.25$.

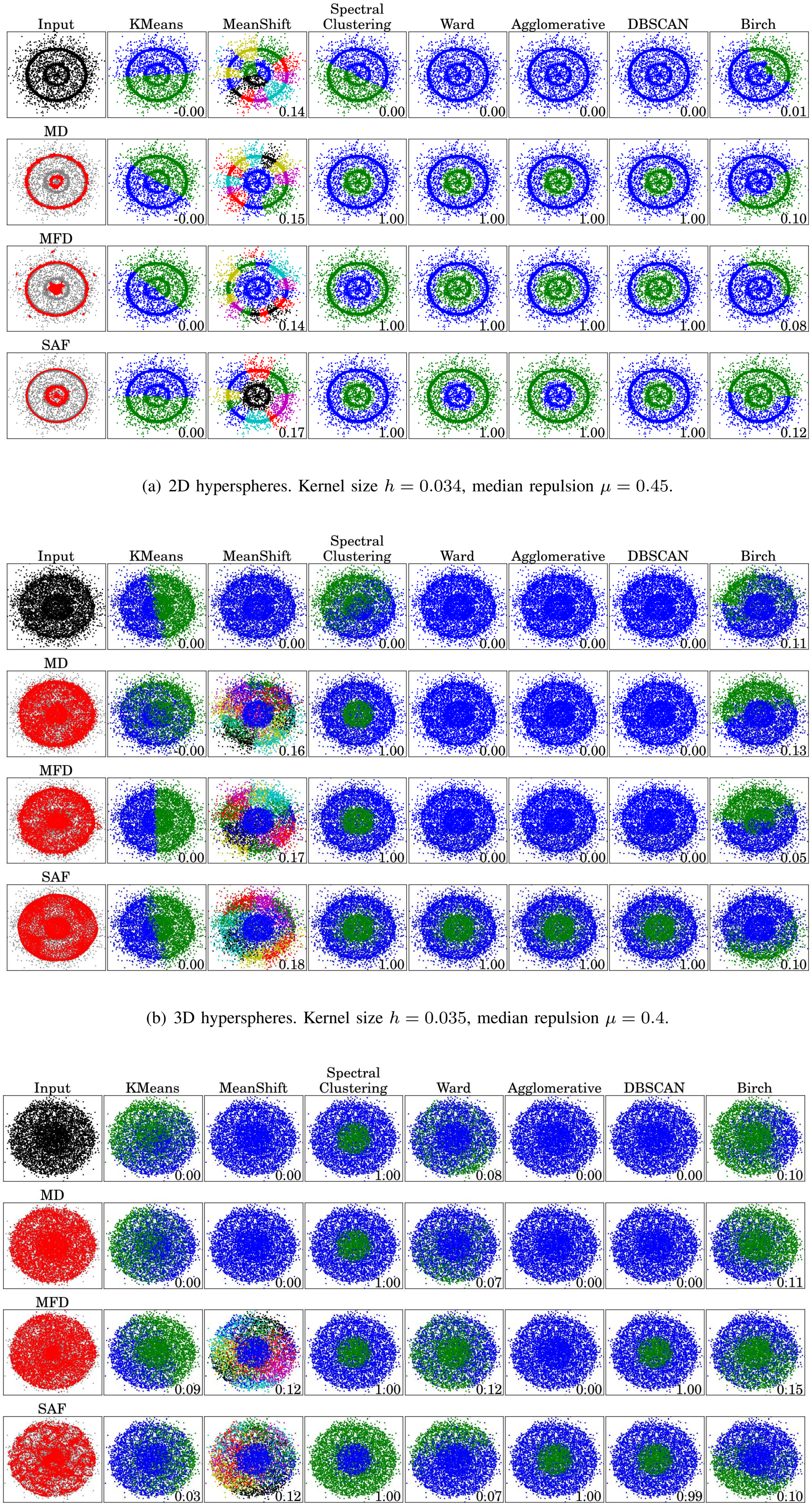
Fig. 8. Additional clustering examples of low dimensional structures corrupted with high dimensional noise. Top row: clustering without consolidation; bottom row: with consolidation. Leftmost column: consolidated data in red. Here the kernel size h is relative to bounding box diagonal. The consolidation outputs are projected to 3D via PCA, followed by clustering in 3D.



(a) 6D to 2D. Gaussian noise with $\sigma = 0.07$, kernel size $h = 0.04$ (relative to bounding box diagonal), median repulsion $\mu = 0.35$.

(b) 10D to 2D. Gaussian noise with $\sigma = 0.062$, kernel size $h = 0.049$ (relative to bounding box diagonal), median repulsion $\mu = 0.4$.

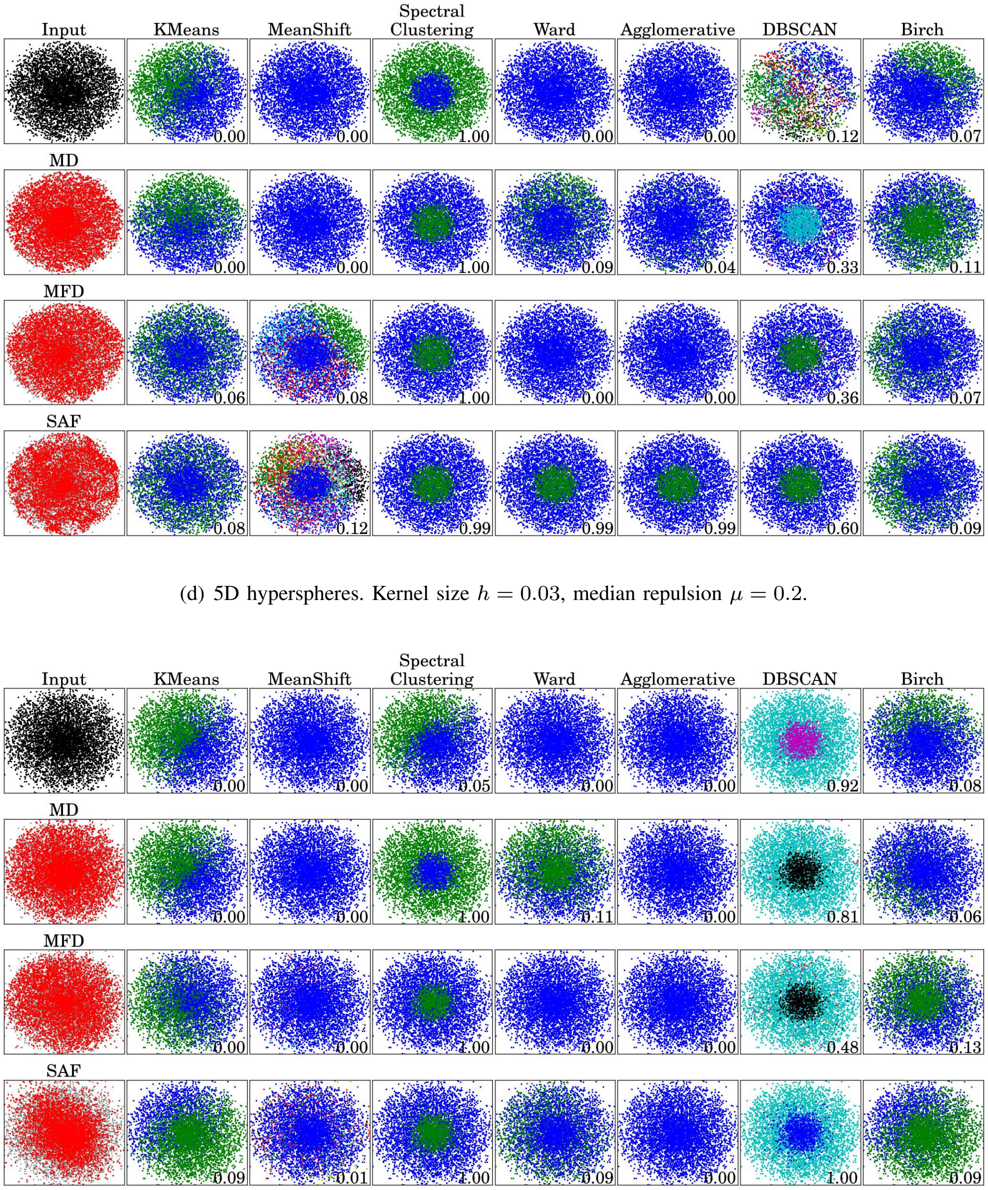
Fig. 9. Performance of different dimensionality reduction techniques without (top rows) and with (bottom rows) SAF consolidation.



(a) 2D hyperspheres. Kernel size $h = 0.034$, median repulsion $\mu = 0.45$.

(b) 3D hyperspheres. Kernel size $h = 0.035$, median repulsion $\mu = 0.4$.

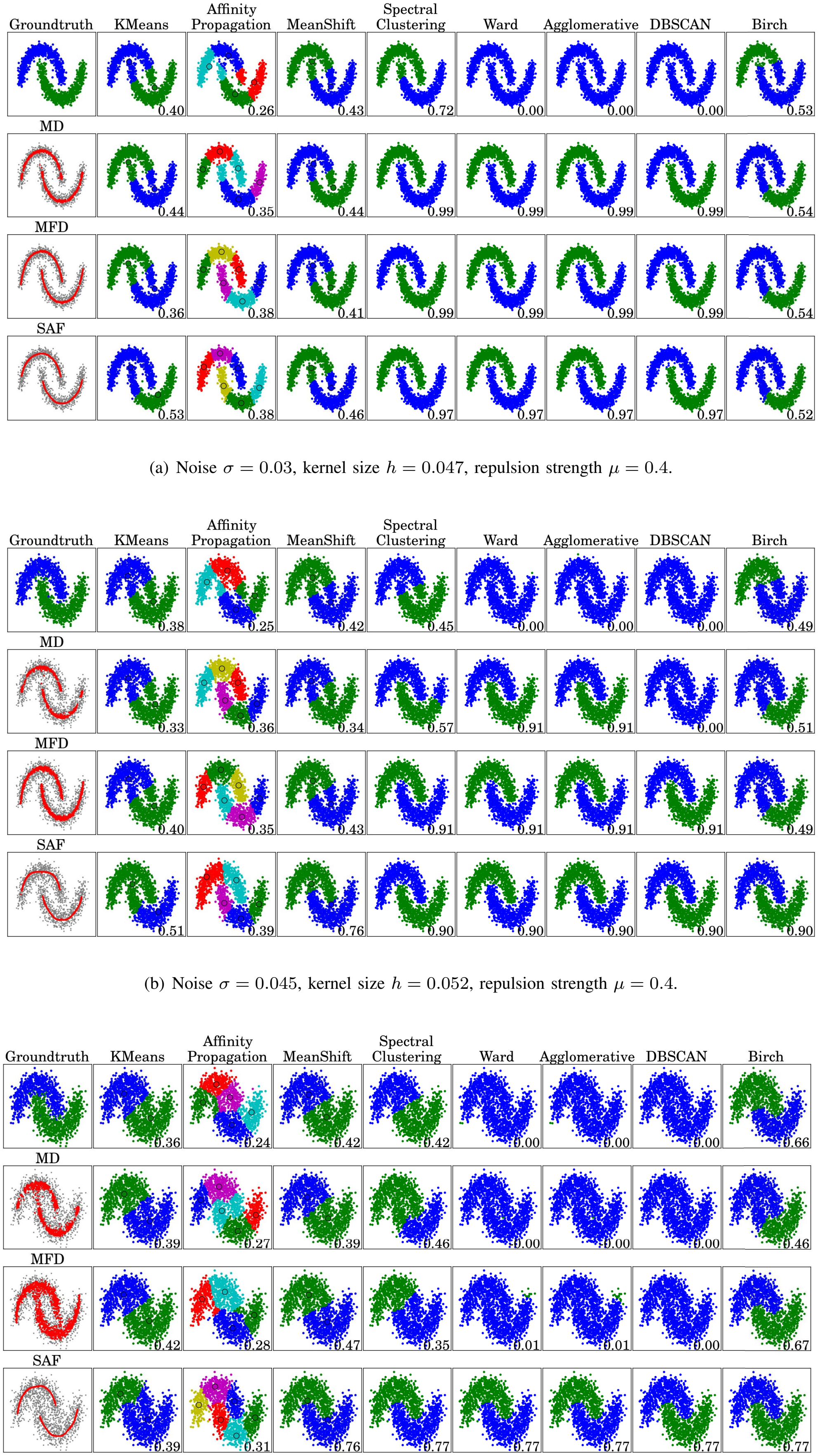
(c) 4D hyperspheres. Kernel size $h = 0.024$, median repulsion $\mu = 0.35$.



(d) 5D hyperspheres. Kernel size $h = 0.03$, median repulsion $\mu = 0.2$.

(e) 6D hyperspheres. Kernel size $h = 0.011$, median repulsion $\mu = 0.15$.

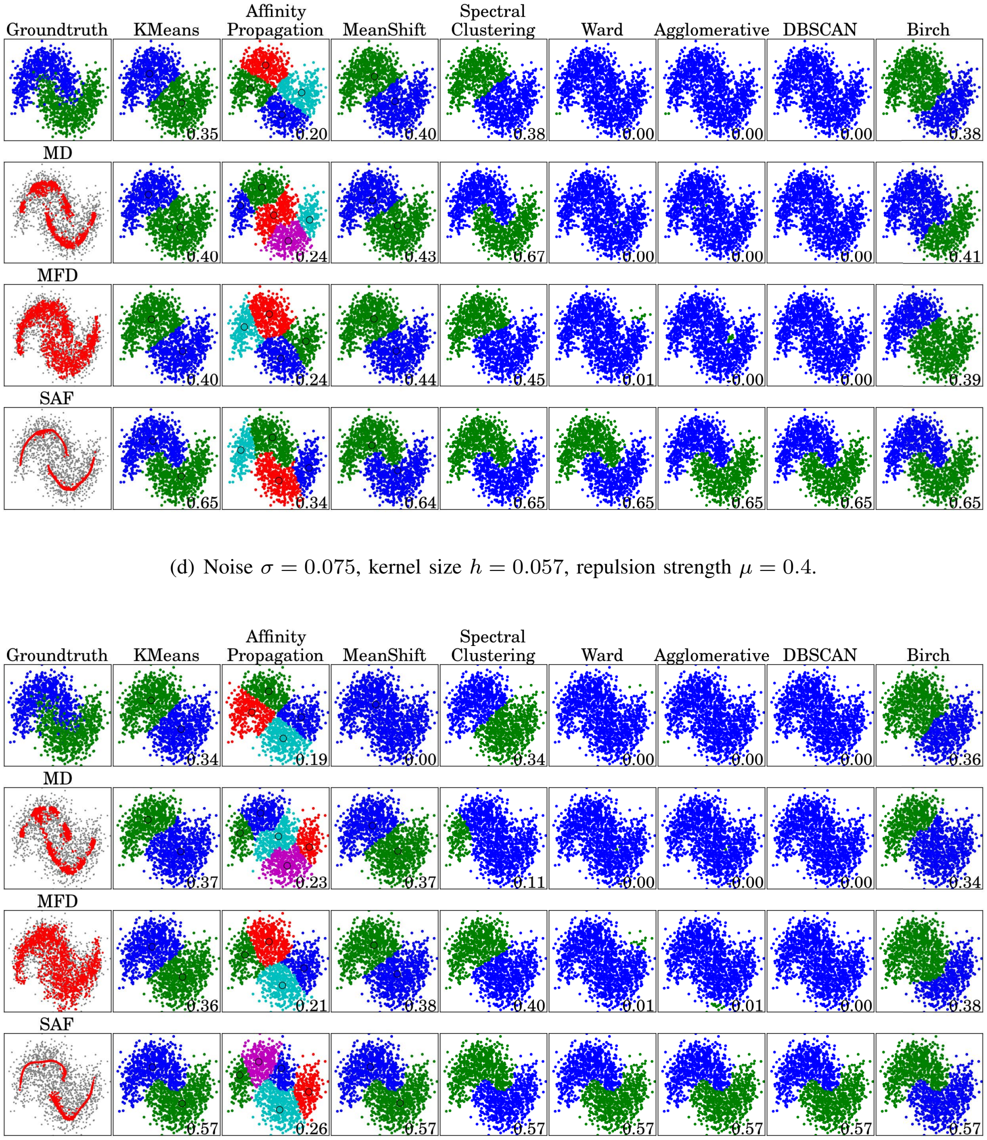
Fig. 10. SAF performance with increasing dimensionality of the data. The top and bottom rows show clustering without and with consolidation, respectively. SAF improves the performance of approaches including spectral clustering, Ward, agglomerative, and DBSCAN as shown, but its effectiveness decreases with data dimensionality increasing. We always use Gaussian noise with $\sigma = 0.05$ (relative to bounding box diagonal), and cluster without dimensionality reduction.



(a) Noise $\sigma = 0.03$, kernel size $h = 0.047$, repulsion strength $\mu = 0.4$.

(b) Noise $\sigma = 0.045$, kernel size $h = 0.052$, repulsion strength $\mu = 0.4$.

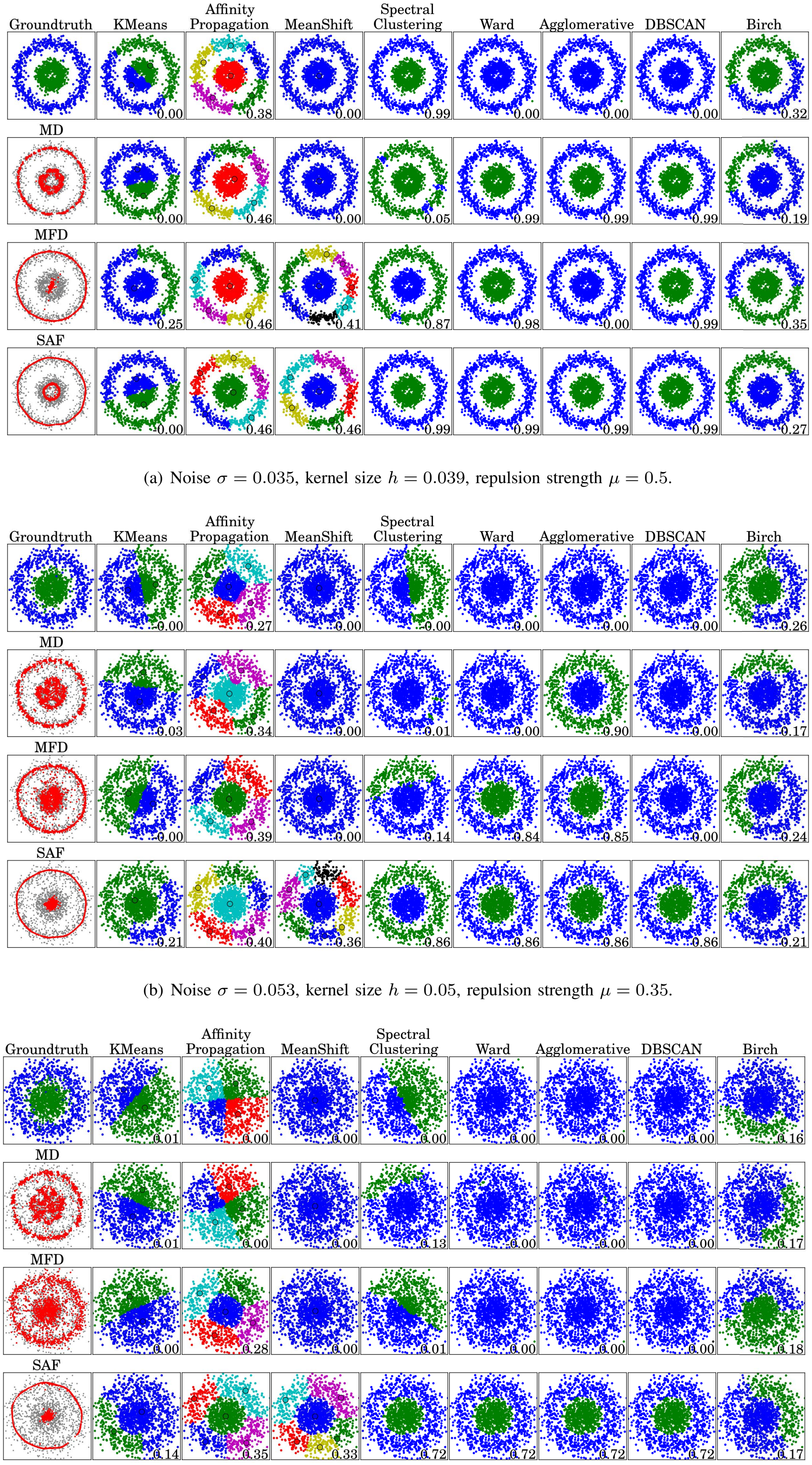
(c) Noise $\sigma = 0.06$, kernel size $h = 0.053$, repulsion strength $\mu = 0.35$.



(d) Noise $\sigma = 0.075$, kernel size $h = 0.057$, repulsion strength $\mu = 0.4$.

(e) Noise $\sigma = 0.09$, kernel size $h = 0.073$, repulsion strength $\mu = 0.35$.

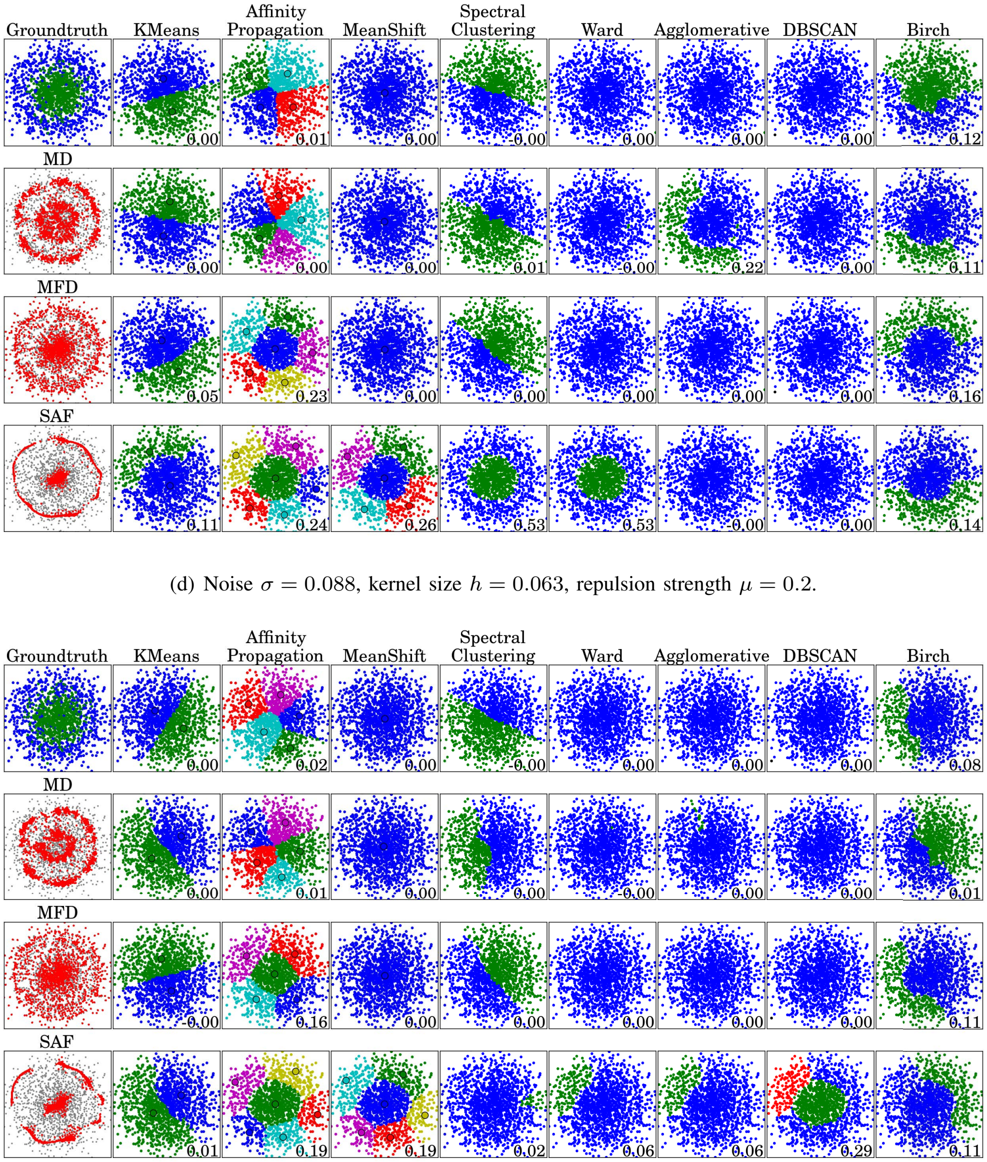
Fig. 11. We evaluate our approach for different noise levels using 2D data. Top row: clustering without SAF consolidation; bottom row: with SAF consolidation. We use anisotropic SAF and cluster in 2D. All σ and h values relative to bounding box diagonal.



(a) Noise $\sigma = 0.035$, kernel size $h = 0.039$, repulsion strength $\mu = 0.5$.

(b) Noise $\sigma = 0.053$, kernel size $h = 0.05$, repulsion strength $\mu = 0.35$.

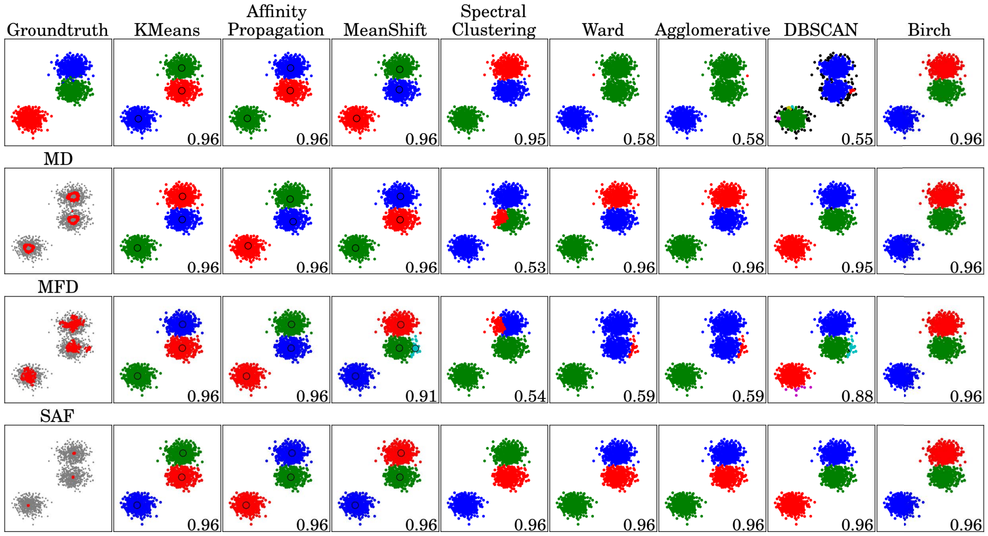
(c) Noise $\sigma = 0.071$, kernel size $h = 0.056$, repulsion strength $\mu = 0.25$.



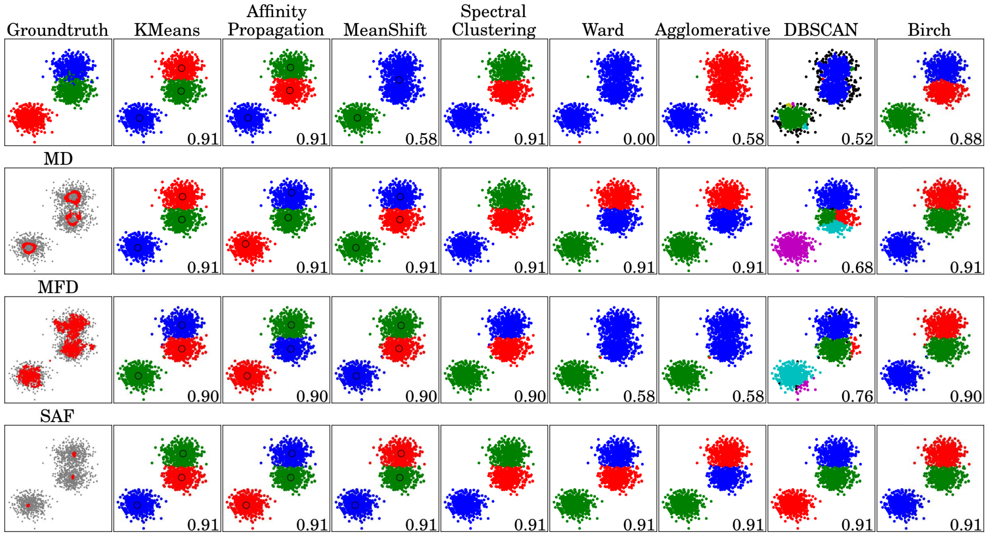
(d) Noise $\sigma = 0.088$, kernel size $h = 0.063$, repulsion strength $\mu = 0.2$.

(e) Noise $\sigma = 0.11$, kernel size $h = 0.069$, repulsion strength $\mu = 0.15$.

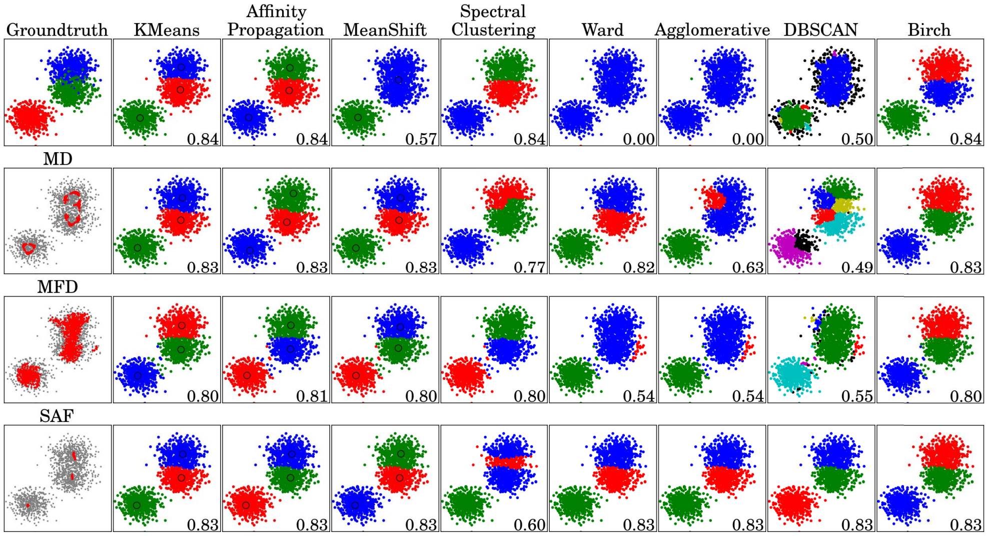
Fig. 12. Comparison of different noise levels using 2D concentric circles, similar to Figure 11.



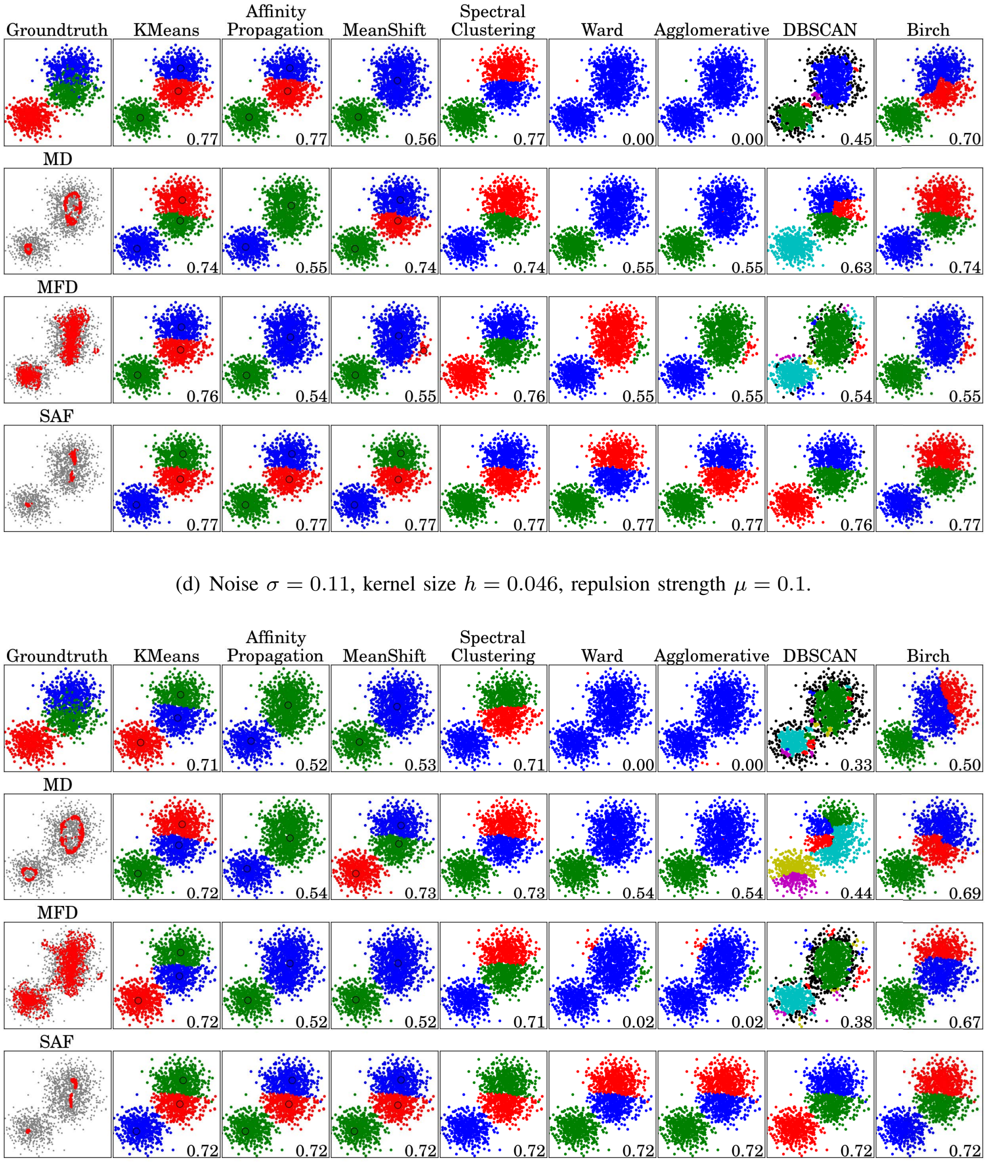
(a) Noise $\sigma = 0.063$, kernel size $h = 0.048$, repulsion strength $\mu = 0.3$.



(b) Noise $\sigma = 0.079$, kernel size $h = 0.04$, repulsion strength $\mu = 0.2$.



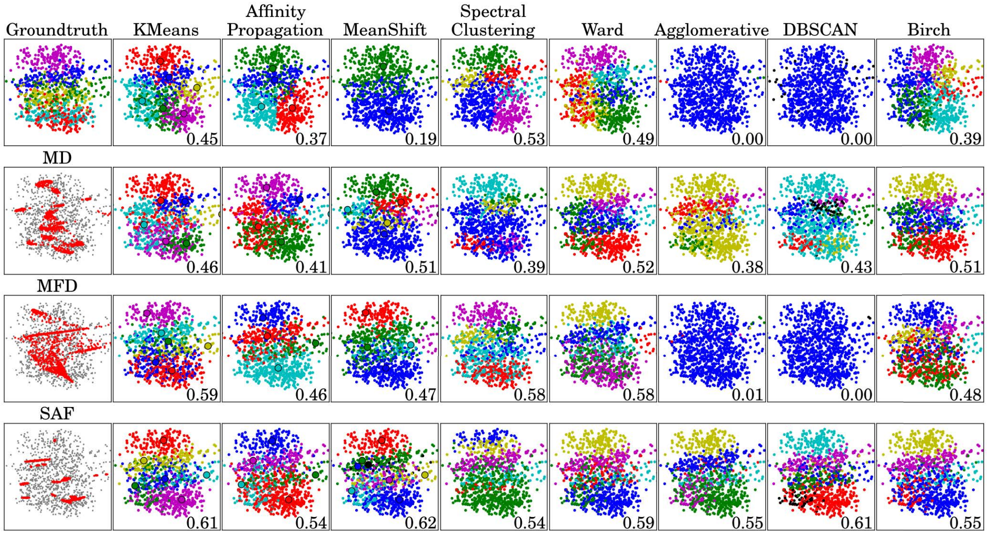
(c) Noise $\sigma = 0.094$, kernel size $h = 0.045$, repulsion strength $\mu = 0.2$.



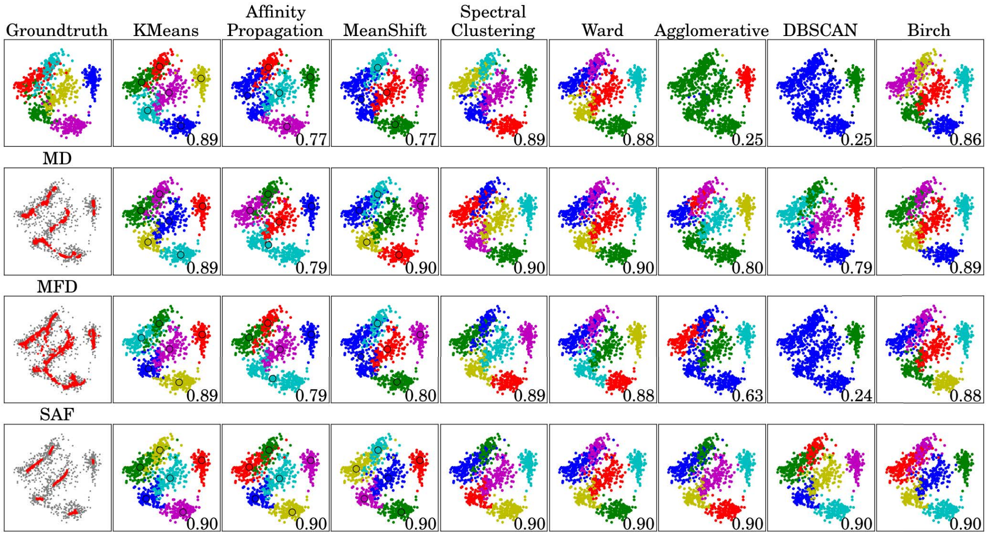
(d) Noise $\sigma = 0.11$, kernel size $h = 0.046$, repulsion strength $\mu = 0.1$.

(e) Noise $\sigma = 0.13$, kernel size $h = 0.047$, repulsion strength $\mu = 0.05$.

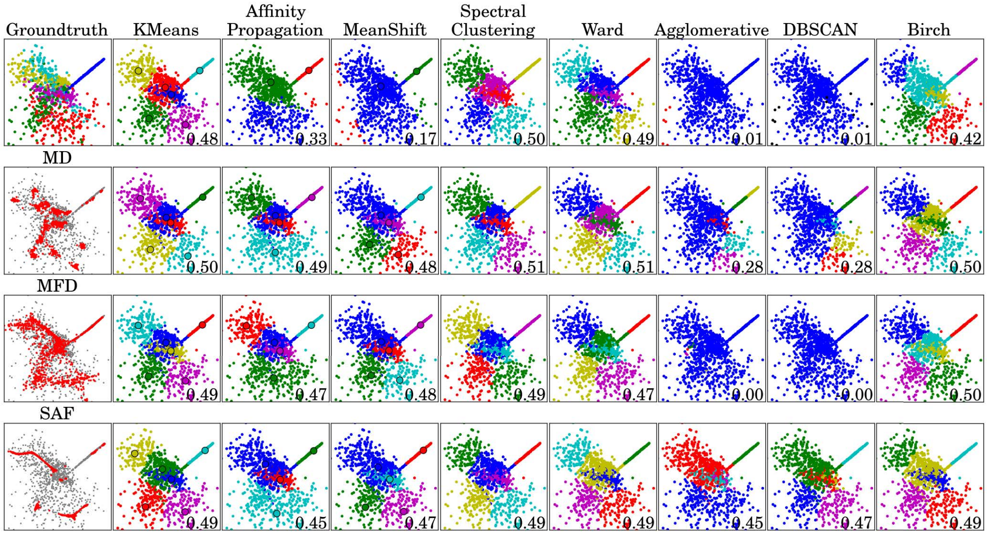
Fig. 13. Comparison of different noise levels using 2D blobs, similar to Figure 11.



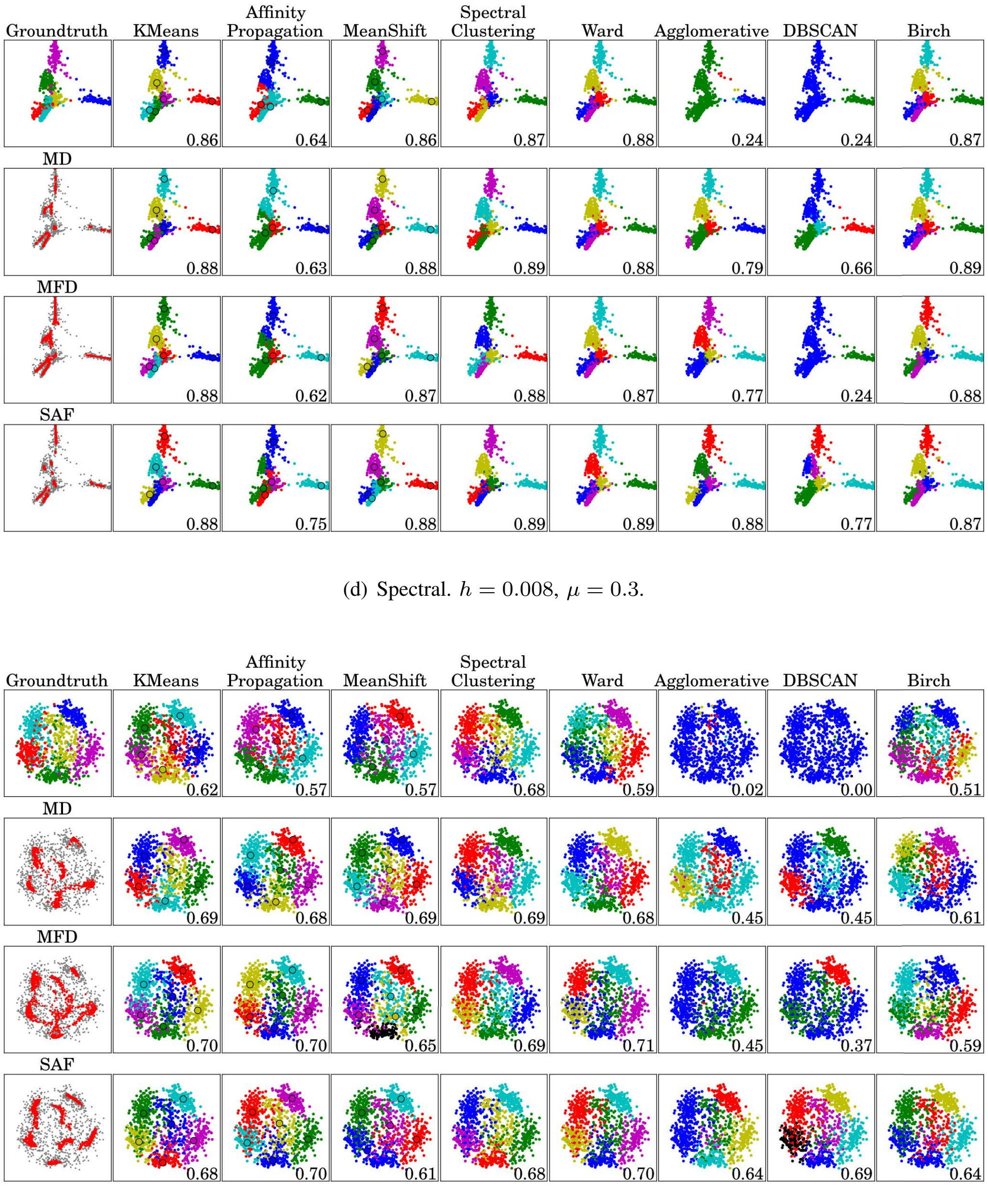
(a) PCA. $h = 0.035$, $\mu = 0.3$.



(b) ISO. $h = 0.022$, $\mu = 0.45$.



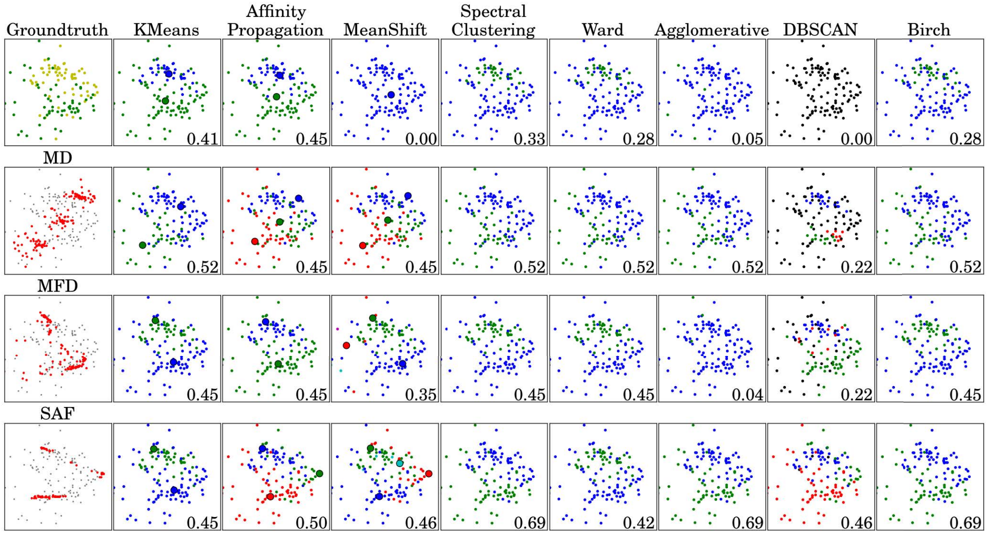
(c) LLE. $h = 0.015$, $\mu = 0.45$.



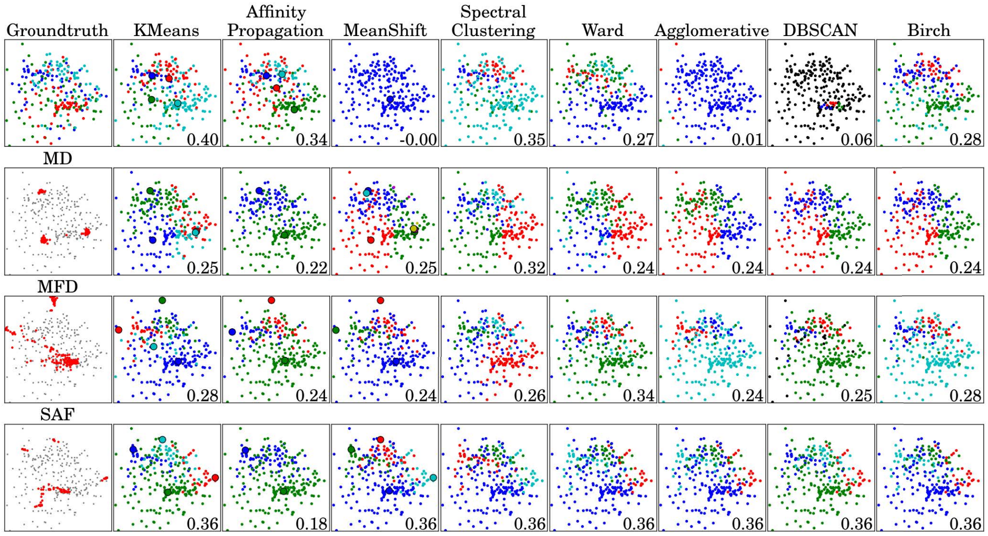
(d) Spectral. $h = 0.008$, $\mu = 0.3$.

(e) MDS. $h = 0.054$, $\mu = 0.2$.

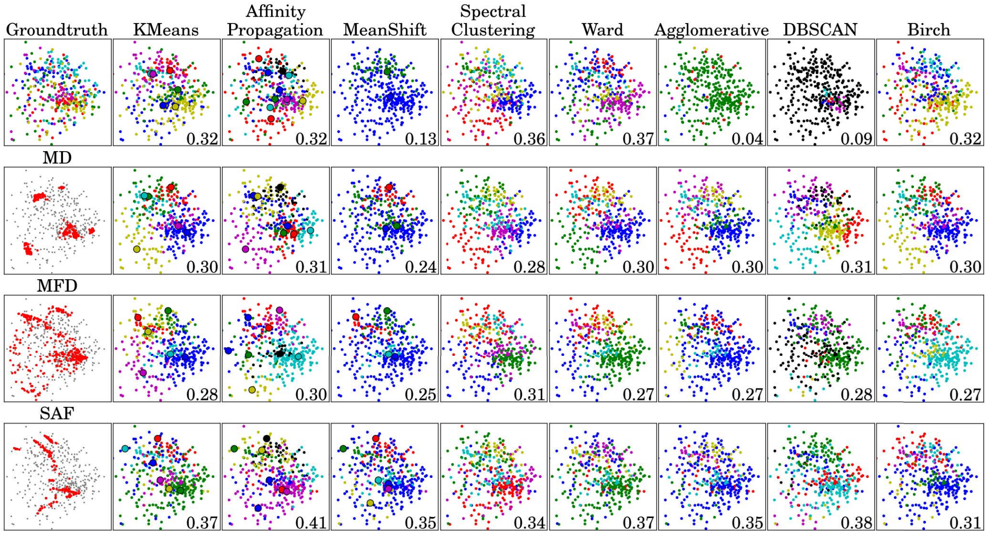
Fig. 14. Clustering the real MINST data with different 3D embedding spaces. Top row: clustering without SAF consolidation; bottom row: clustering with SAF consolidation. All h values relative to bounding box diagonal. We perform SAF in 3D before clustering.



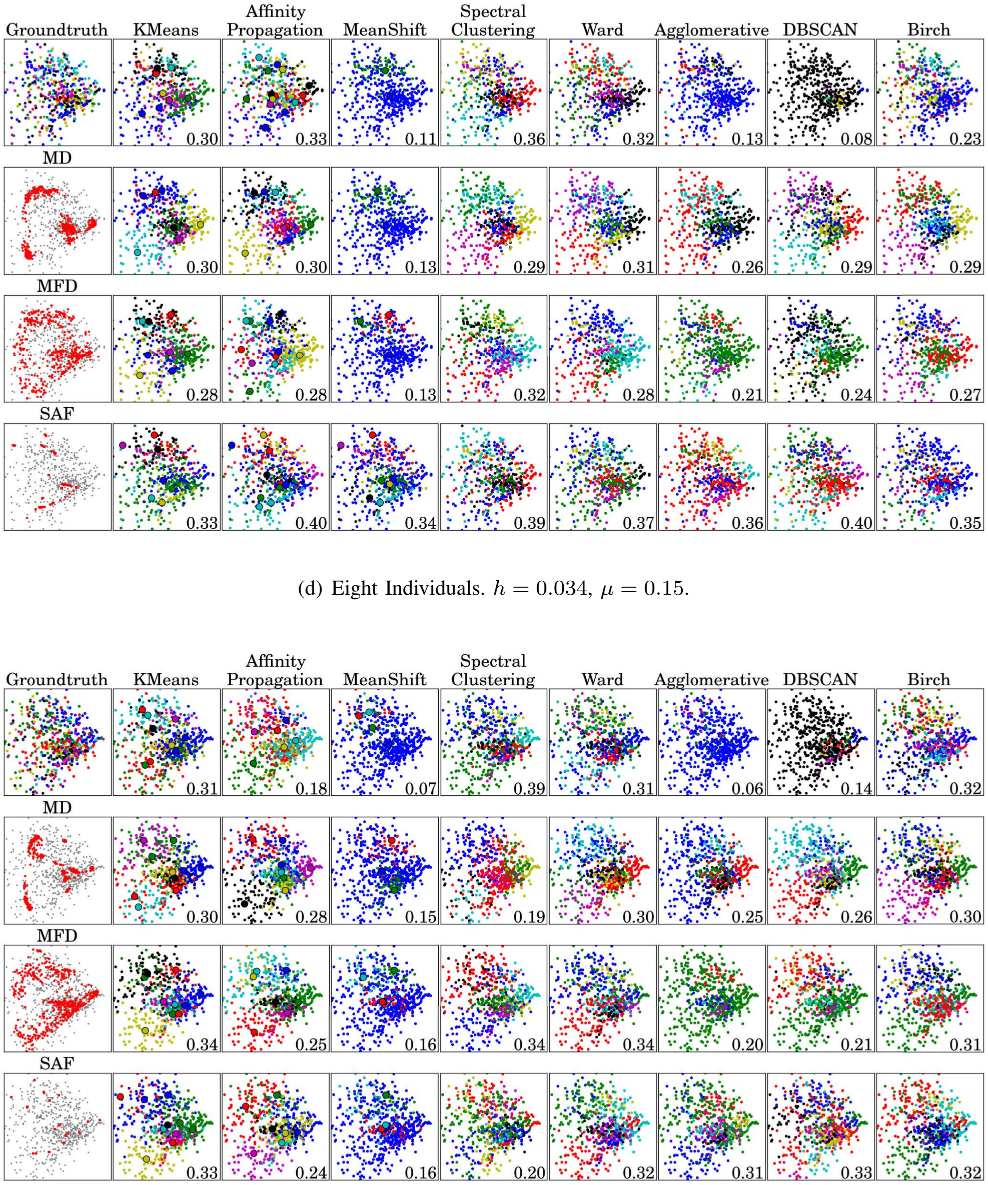
(a) Two Individuals. $h = 0.06$, $\mu = 0.25$.



(b) Four Individuals. $h = 0.047$, $\mu = 0.15$.



(c) Six Individuals. $h = 0.036$, $\mu = 0.15$.



(d) Eight Individuals. $h = 0.034$, $\mu = 0.15$.

(e) Ten Individuals. $h = 0.033$, $\mu = 0.1$.

Fig. 15. We illustrate the performance of our consolidation with increasing number of target clusters in the Yale face data. Top row: clustering without SAF consolidation; bottom row: clustering after SAF consolidation. All h values are relative to the bounding box diagonal of the data. The data is pre-processed before SAF as described in the paper (Section 4.2.5).