Clustered Sparse Bayesian Learning: Supplementary File

Yu Wang [*] **David Wipf** [†] **Jeong-Min Yun** [♯] **Wei Chen**[*] **Ian Wassell** [*]

[*] University of Cambridge, Cambridge, UK
[†] Microsoft Research, Beijing, China
[♯]Pohang University of Science and Technology, Pohang, Republic of Korea

1 C-SBL Algorithm Summary

Algorithm 1 Clustered Sparse Bayesian Learning Algorithm (C-SBL)

```
Input: Sensing matrices $\Phi_j,\ j = 1...L$, measurement matrix $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, ...\mathbf{y}_L]$, $\beta \geq 0$.
Initialize: $W$, $\Lambda_k\ \forall k$, and $\nu$.
for halting criterion false do
    $\Gamma_j \leftarrow \left[\sum_k w_{j,k} \Lambda_k^{-1}\right]^{-1}, \quad \forall j.$
    $\boldsymbol{z}_j \leftarrow \text{diag}\left[\left(\sum_k w_{j,k} \Lambda_k^{-1} + \frac{1}{\nu}\Phi_j^T \Phi_j\right)^{-1}\right], \quad \forall j.$
    $\boldsymbol{x}_j \leftarrow \Gamma_j \Phi_j^T \left(\nu I + \Phi_j \Gamma_j \Phi_j^T\right)^{-1} \boldsymbol{y}_j, \quad \forall j.$
    $\lambda_{i,k} \leftarrow \frac{\sum_j w_{j,k}\left(x_{i,j}^2 + z_{i,j}\right)}{\sum_j w_{j,k}}, \quad \forall i, k.$
    $w_{j,k} \leftarrow \exp\left(\frac{1}{\beta}\left[-\sum_{i=1}^M \left(\frac{x_{i,j}^2}{\lambda_{i,k}} + \log \lambda_{i,k} + \frac{z_{i,j}}{\lambda_{i,k}}\right)\right] - 1\right), \quad \forall j, k.$
    $w_{j,k} \leftarrow \frac{w_{j,k}}{\sum_k w_{j,k}}, \quad \forall j, k.$
    $\nu \leftarrow \frac{1}{LN}\sum_{j=1}^L \left[\text{tr}\left(\frac{1}{\nu} I + (\Phi_j \Gamma_j \Phi_j^T)^{-1}\right) + \|\boldsymbol{y}_j - \mathbf{\Phi}_j \boldsymbol{x}_j\|_2^2\right].$
end for
return
```

[*]Y. Wang is sponsored by the University of Cambridge Overseas Trust. Y. Wang and J. Yun are partially supported by sponsorship from Microsoft Research Asia. W. Chen is supported by EPSRC Research Grant EP/K033700/1 and the NSFC Research Grant 61401018.

2 Full Image Reconstruction Results (64×64 Images)

We present the complete image reconstruction results described in Section 6 of our submission followed by the corresponding clustering information (heat-maps) from the estimated W matrix. Further explanations below.

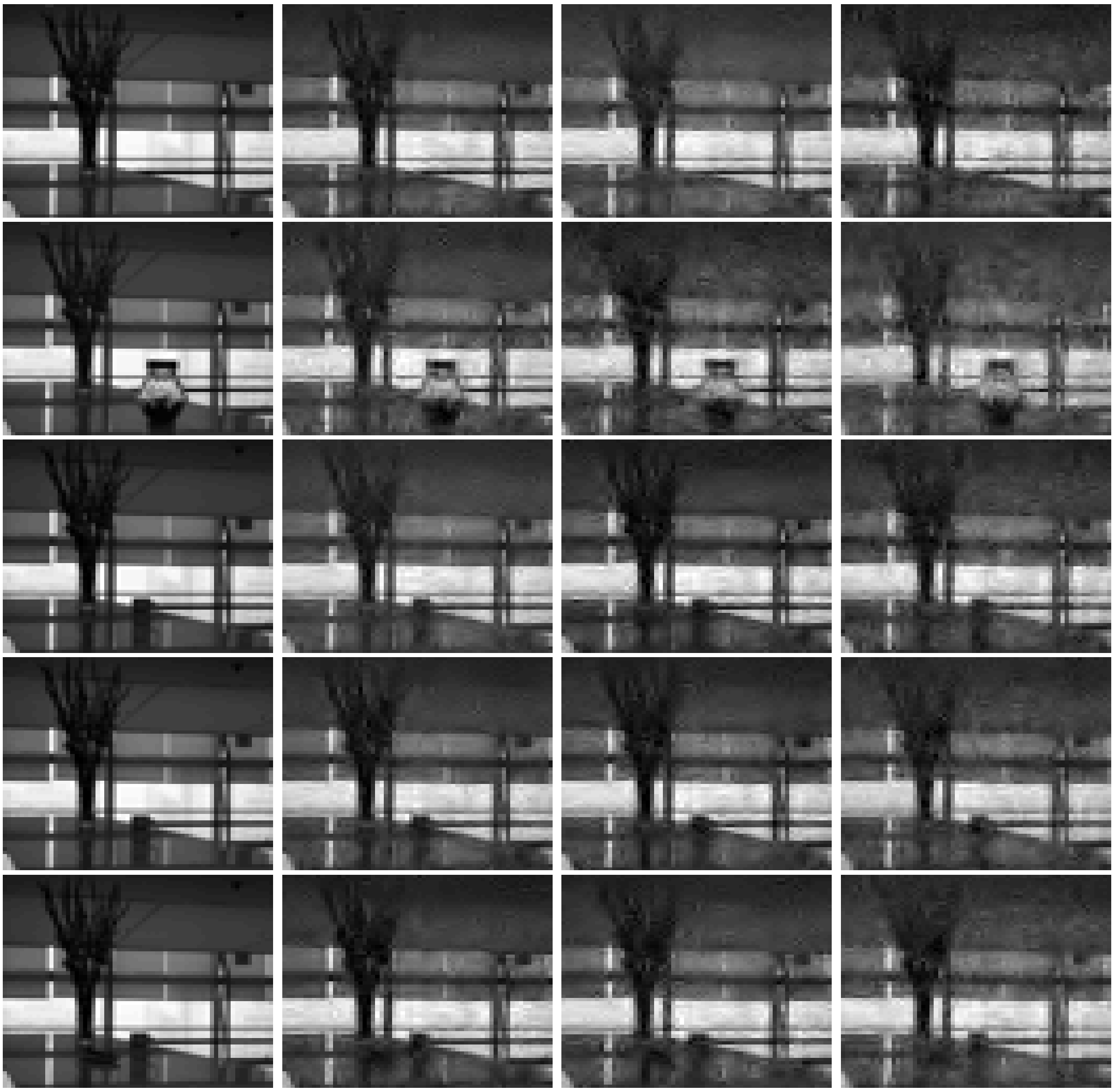


Figure 1: Reconstructions of 64×64 images from dynamic scene 1, cluster size 5. From left to right: Original image, C-SBL, DP, MMV. Sampling rate is $N/M = 0.38$.

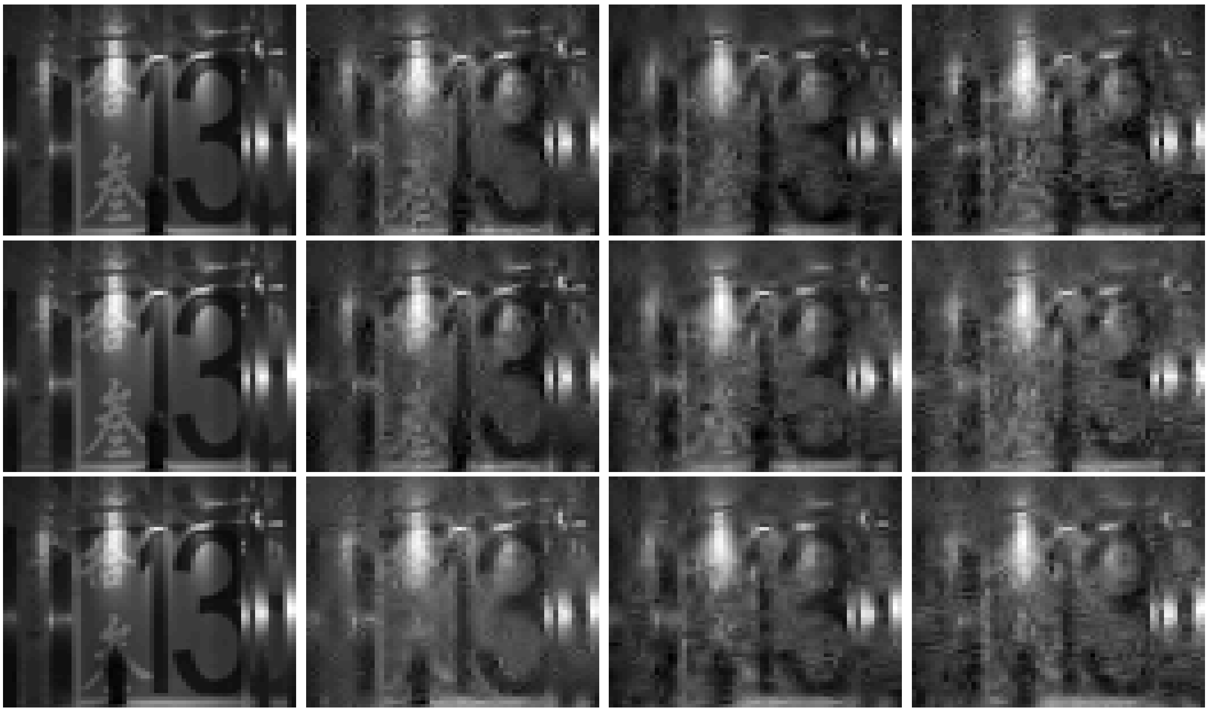


Figure 2: Reconstructions of 64×64 images from dynamic scene 2, cluster size 3. From left to right: Original image, C-SBL, DP, MMV. Sampling rate is $N/M = 0.38$.

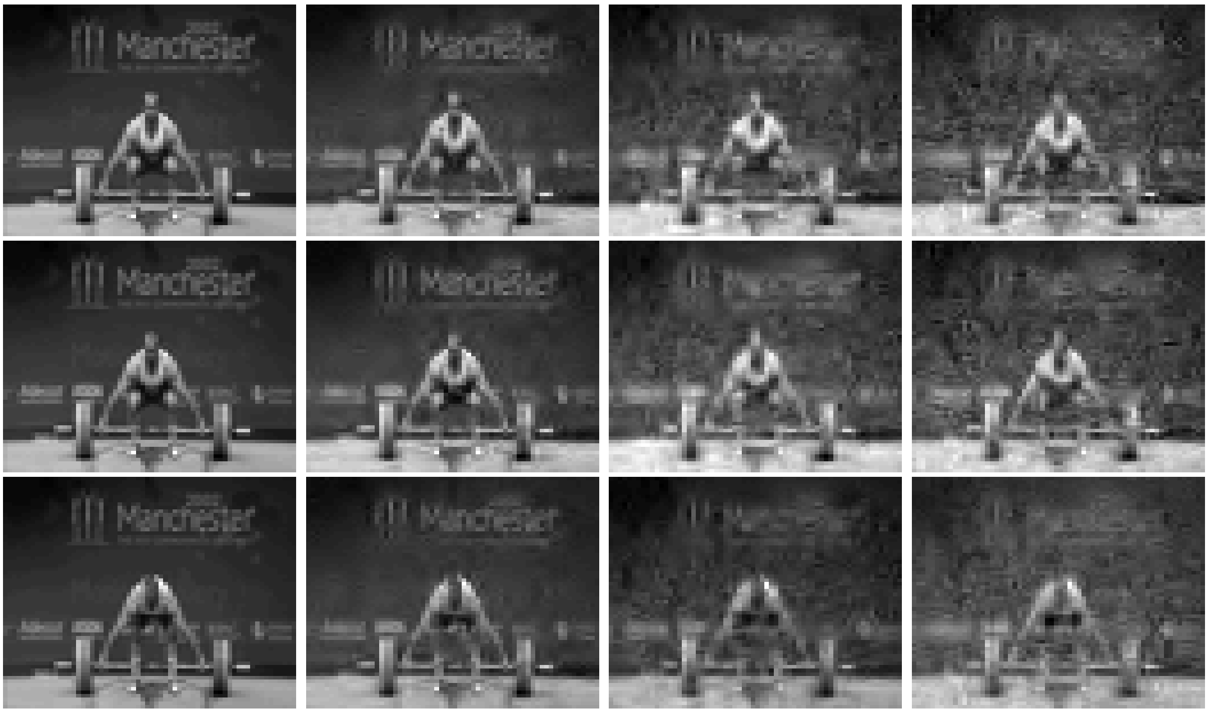


Figure 3: Reconstructions of 64×64 images from dynamic scene 3, cluster size 3. From left to right: Original image, C-SBL, DP, MMV. Sampling rate is $N/M = 0.38$.

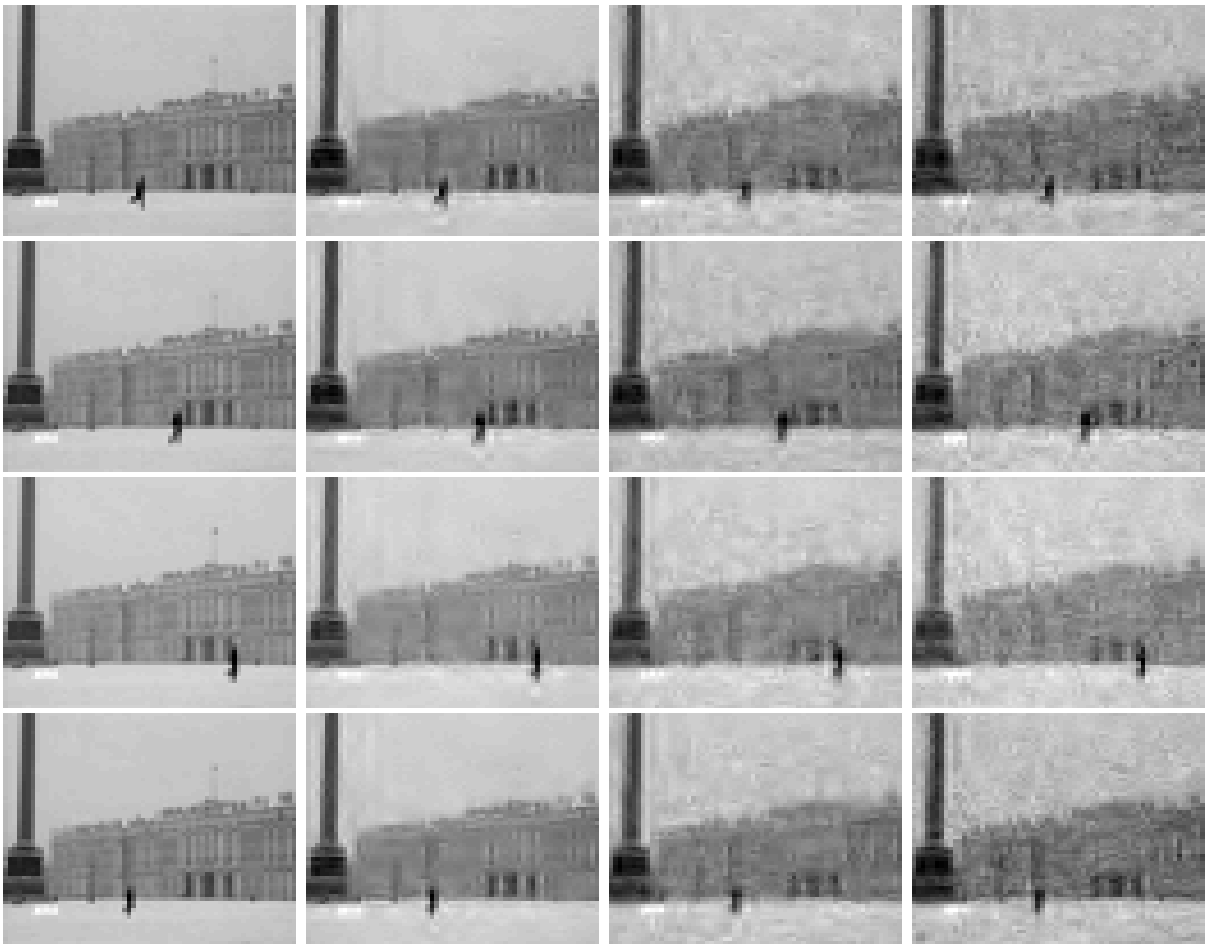


Figure 4: Reconstructions of 64×64 images from dynamic scene 4, cluster size 4. From left to right: Original image, C-SBL, DP, MMV. Sampling rate is $N/M = 0.38$.

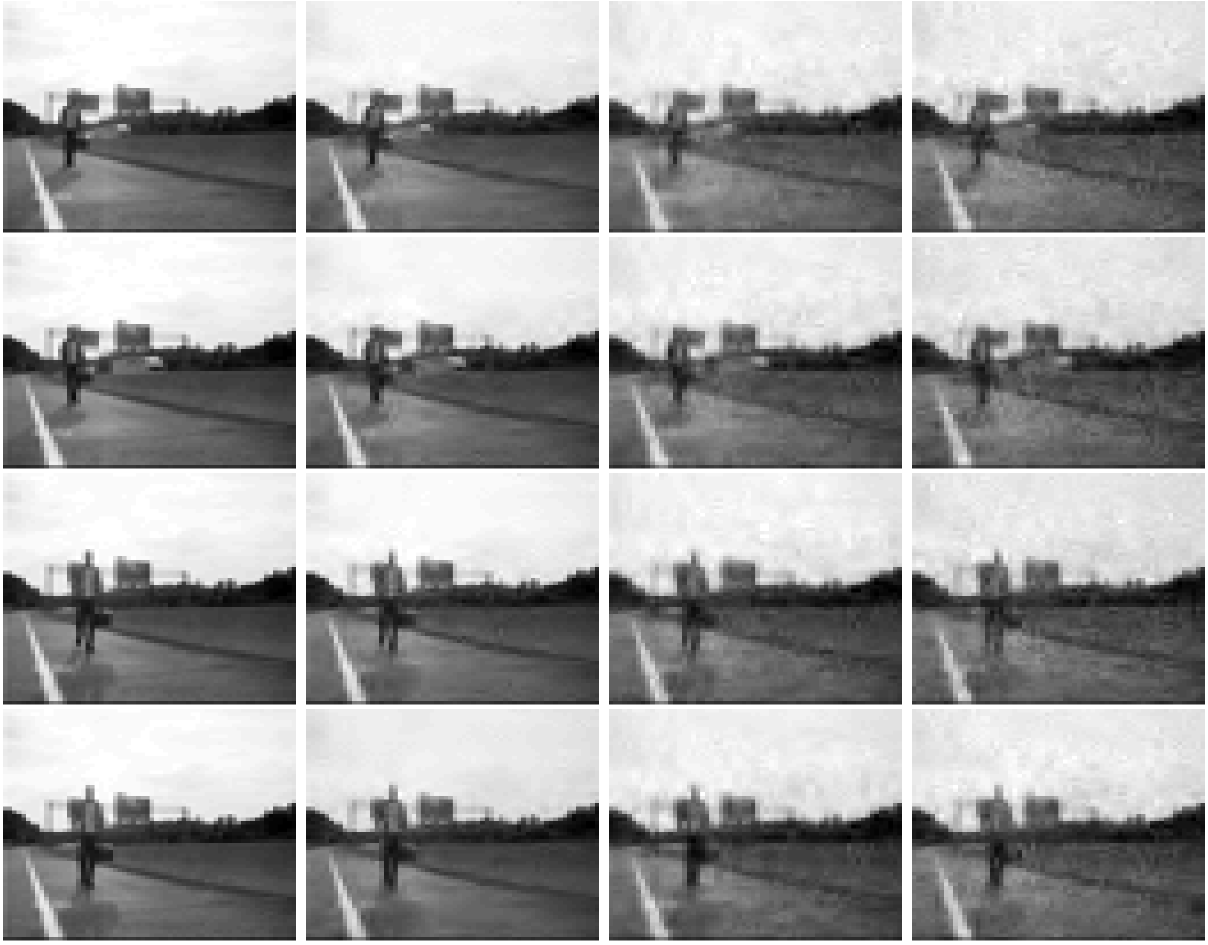
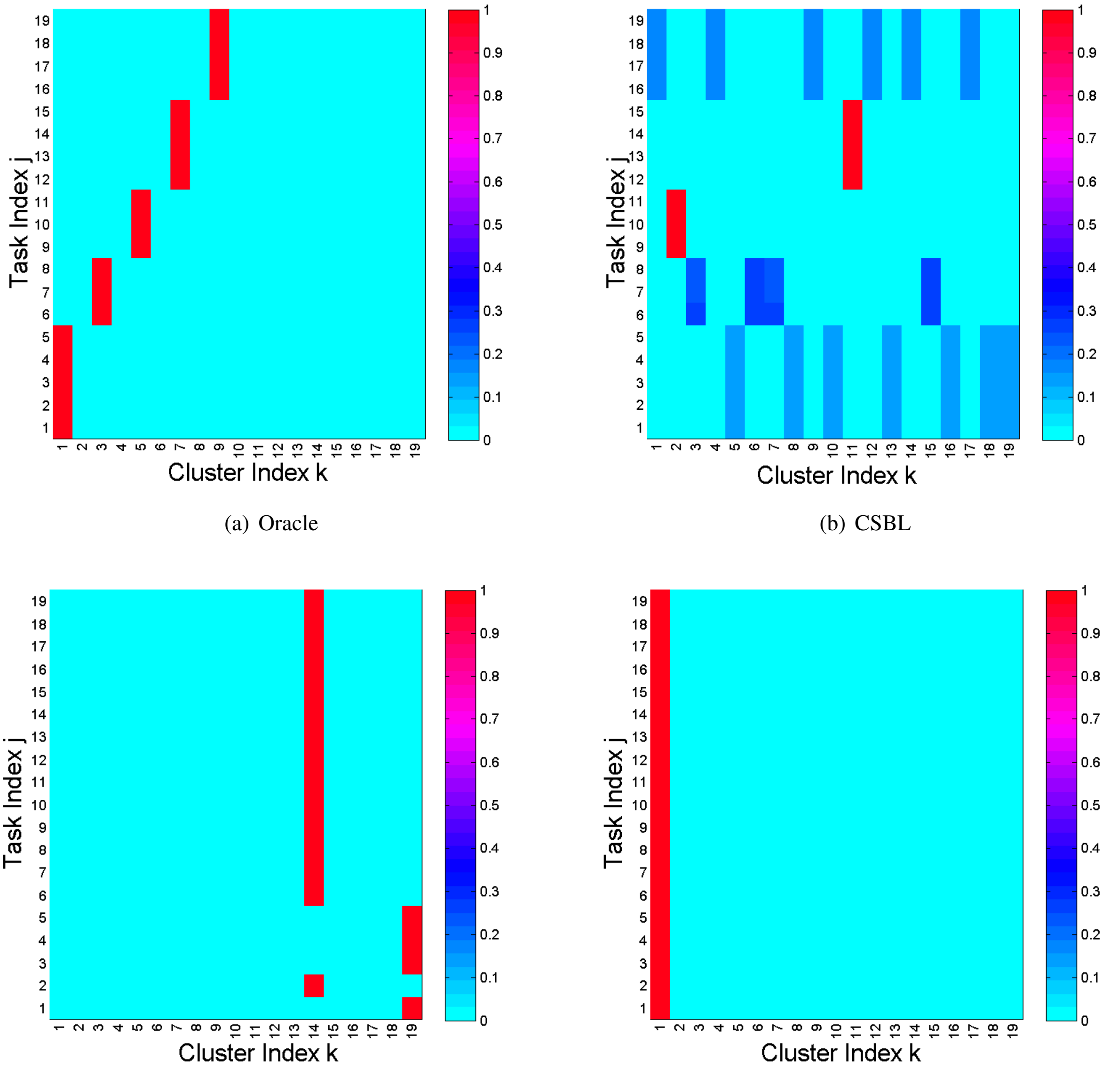


Figure 5: Reconstructions of 64×64 images from dynamic scene 5, cluster size 4. From left to right: Original image, C-SBL, DP, MMV. Sampling rate is $N/M = 0.38$.



(a) Oracle

(b) CSBL

(c) DP

(d) MMV

Figure 6: Heat-map of W matrices for 64×64 image reconstructions. (a) Oracle case (cluster patterns are known), (b) C-SBL, (c) DP, and (d) MMV.

Figure 6 displays the clustering information of the respective algorithms with respect to ground truth as revealed through heat-maps of the estimated cluster matrices W. Here column permutations are irrelevant as the column labels are arbitrary; more important is that tasks within the same group (as partitioned by the oracle) have nonzeros along the same columns of W. Note that because the MMV algorithm assumes a single cluster, it is implicitly associated with the degenerate W matrix shown in the figure. In contrast, DP learns only two clusters, errantly merging many distinct categories together, which is arguably a primary contributor to its reconstruction error. Meanwhile C-SBL correctly learns the five correct clusters and achieves the best performance, both in terms of MSE (as reported in our submission) and visual inspection.

In terms of the underlying representations, unlike DP, C-SBL uses multiple bases Λ_k to model many of the clusters (scenes) as evidenced by multiple nonzeros in the rows of W. However, this is just an artifact of many different Λ_k fusing together within a true cluster, and all of these bases within a cluster must eventually share the same support (and typically magnitudes as well) by virtue of the support intersection property described in our submission. In contrast, DP seems to prematurely allocate its corresponding $w_{j,k}$ values to a single basis per task, eventually becoming trapped at suboptimal extrema.

Finally, we should mention that in certain more challenging problems with sparse innovation components C-SBL can potentially overestimate the number of clusters. However, often unnecessary splits do not actually affect the corresponding reconstruction error because the within cluster row-sparsity estimation of C-SBL is still extremely effective. In other words, while superfluous cluster splits can often be compensated for, unwarranted cluster merging as demonstrated by DP typically cannot be.

3 Higher Resolution Full Image Reconstruction Results (128×128 Images)

This section demonstrates analogous image reconstruction performance at a higher resolution. We sample 128×128 versions of the same 19 images again in the wavelet domain and using 19 distinct sensing matrices $\Phi_j \ , j = 1 ... L$ as before. Because of the higher resolution, the images now become sparser in the wavelet domain. We correspondingly reduce the sampling rate to $N/M = 0.33$ in order to challenge the estimation capability of all the learning approaches. The improvement afforded by C-SBL is more notably apparent in images having text.

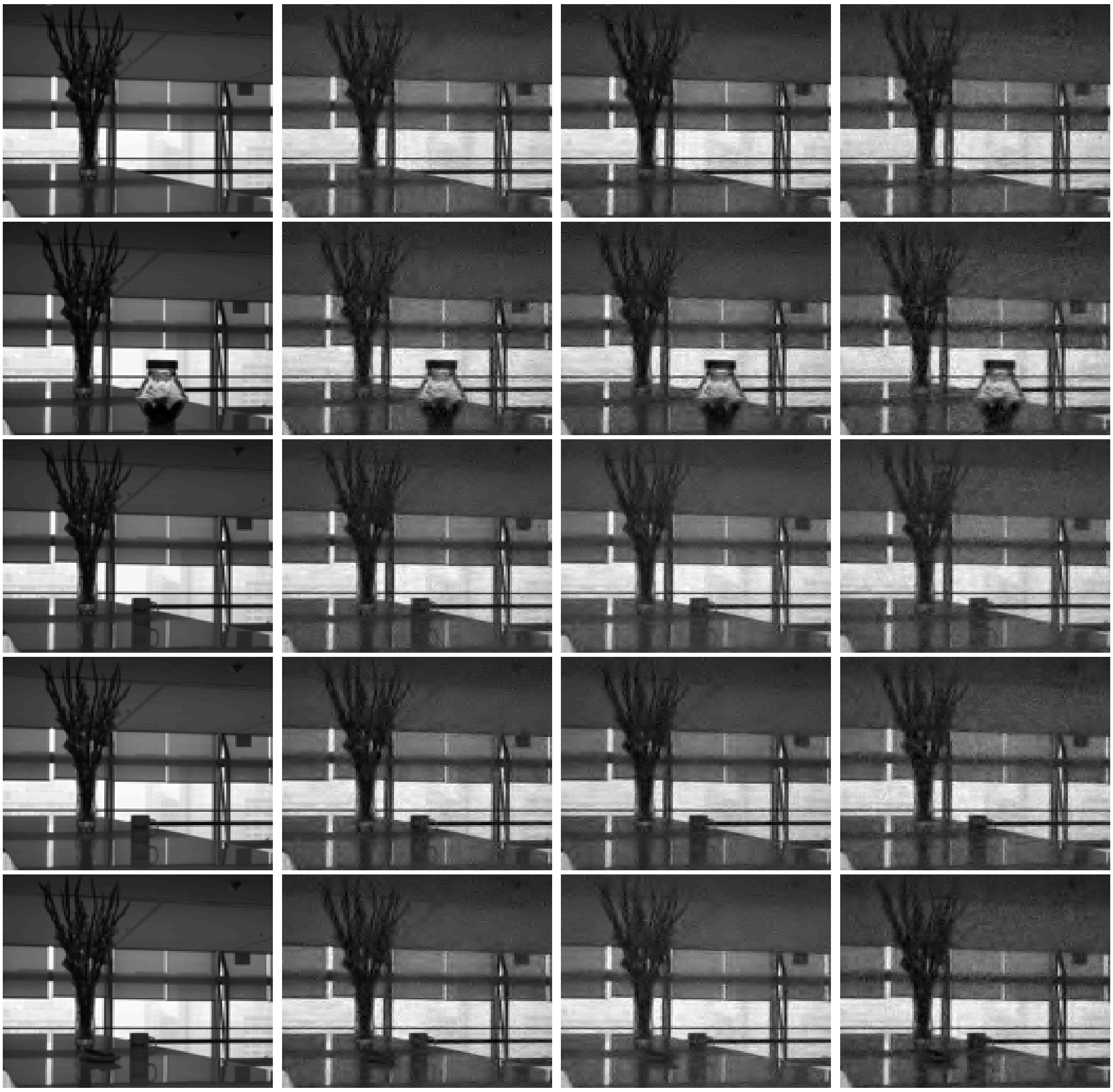


Figure 7: Reconstructions of 128×128 images from dynamic scene 1, cluster size 5. From left to right: Original image, C-SBL, DP, MMV. Sampling rate is $N/M = 0.33$.

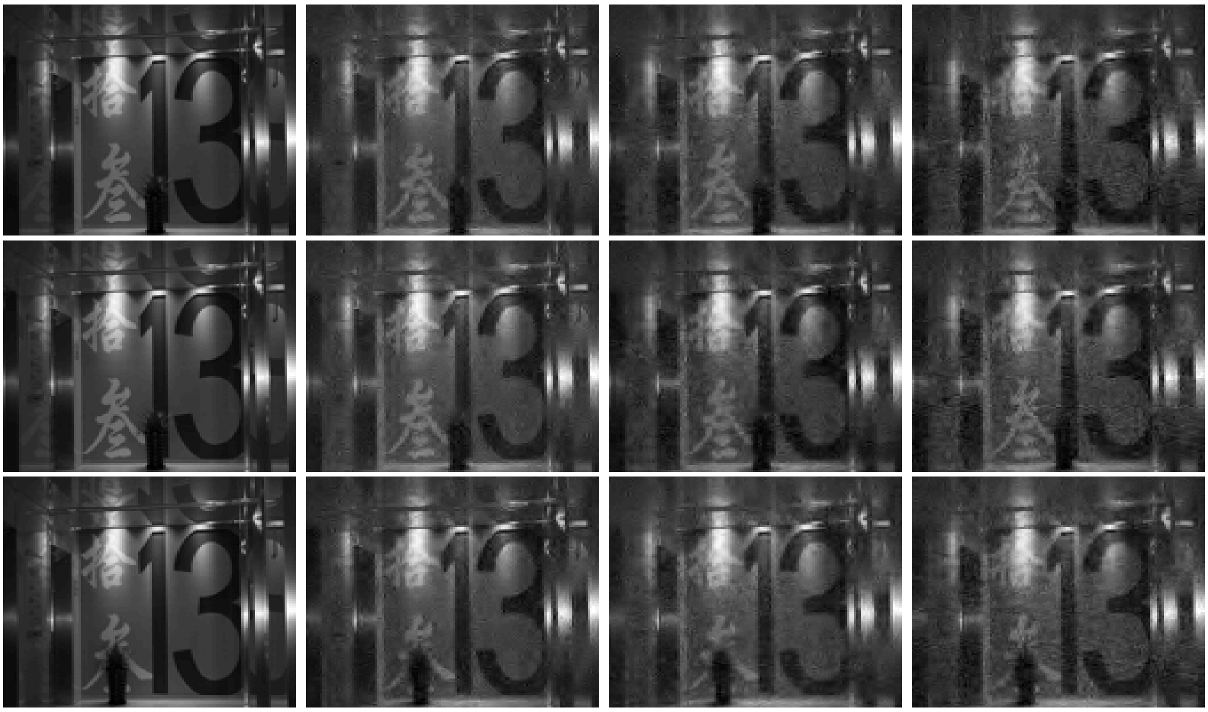


Figure 8: Reconstructions of 128×128 images from dynamic scenes 2, cluster size 3. From left to right: Original image, C-SBL, DP, MMV. Sampling rate is $N/M = 0.33$.

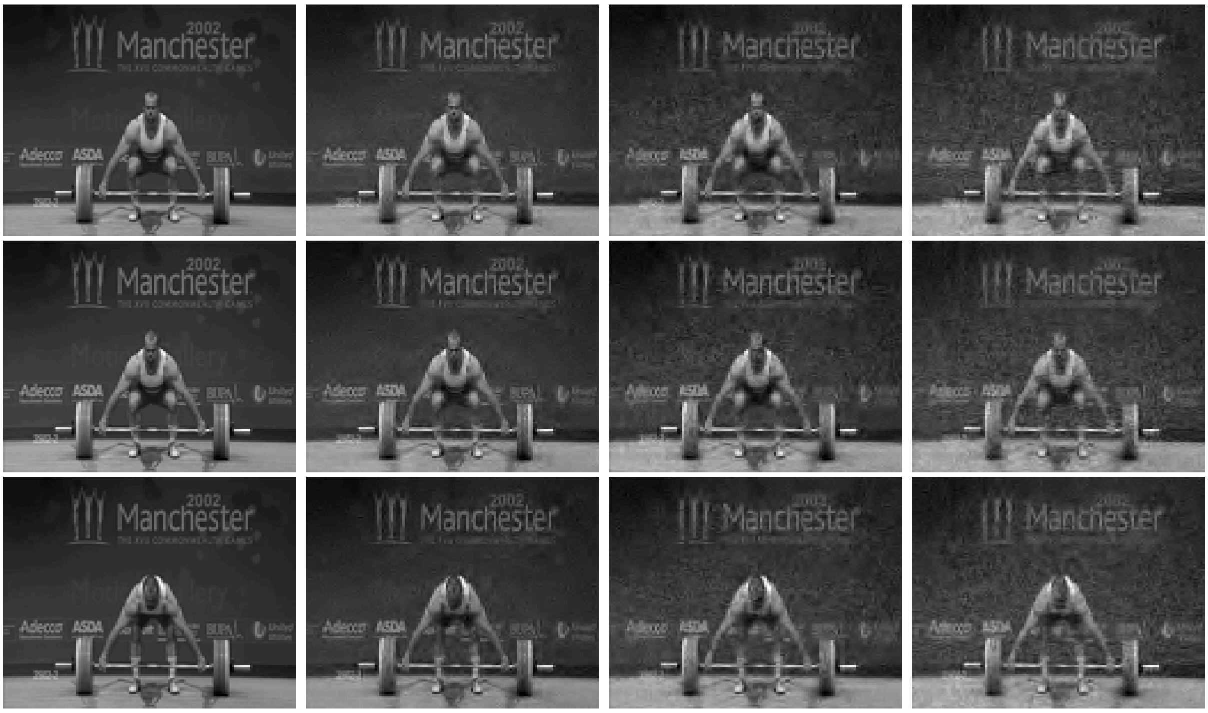


Figure 9: Reconstructions of 128×128 images from dynamic scene 3, cluster size 3. From left to right: Original image, C-SBL, DP, MMV. Sampling rate is $N/M = 0.33$.

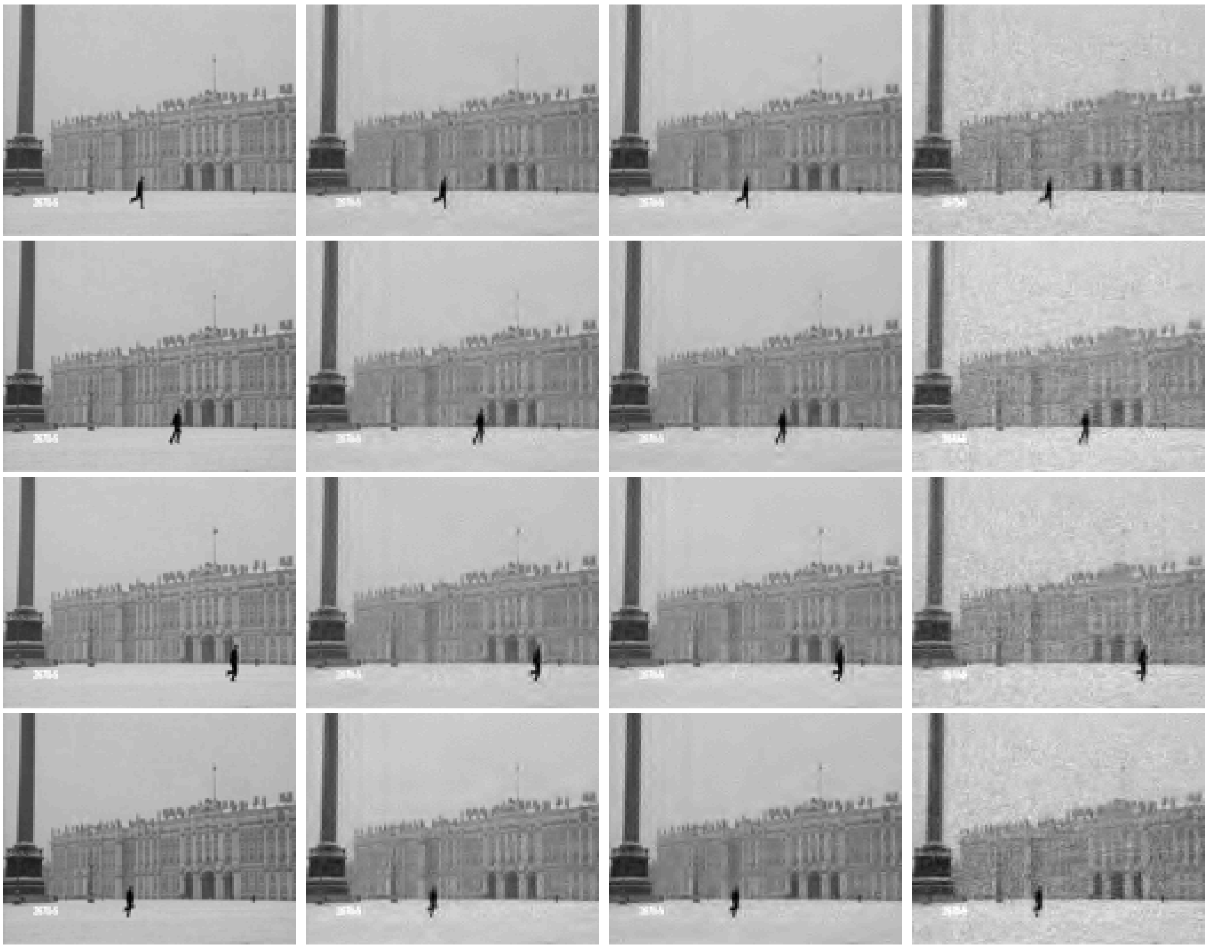


Figure 10: Reconstructions of 128×128 images from dynamic scene 4, cluster size 4. From left to right: Original image, C-SBL, DP, MMV. Sampling rate is $N/M = 0.33$.

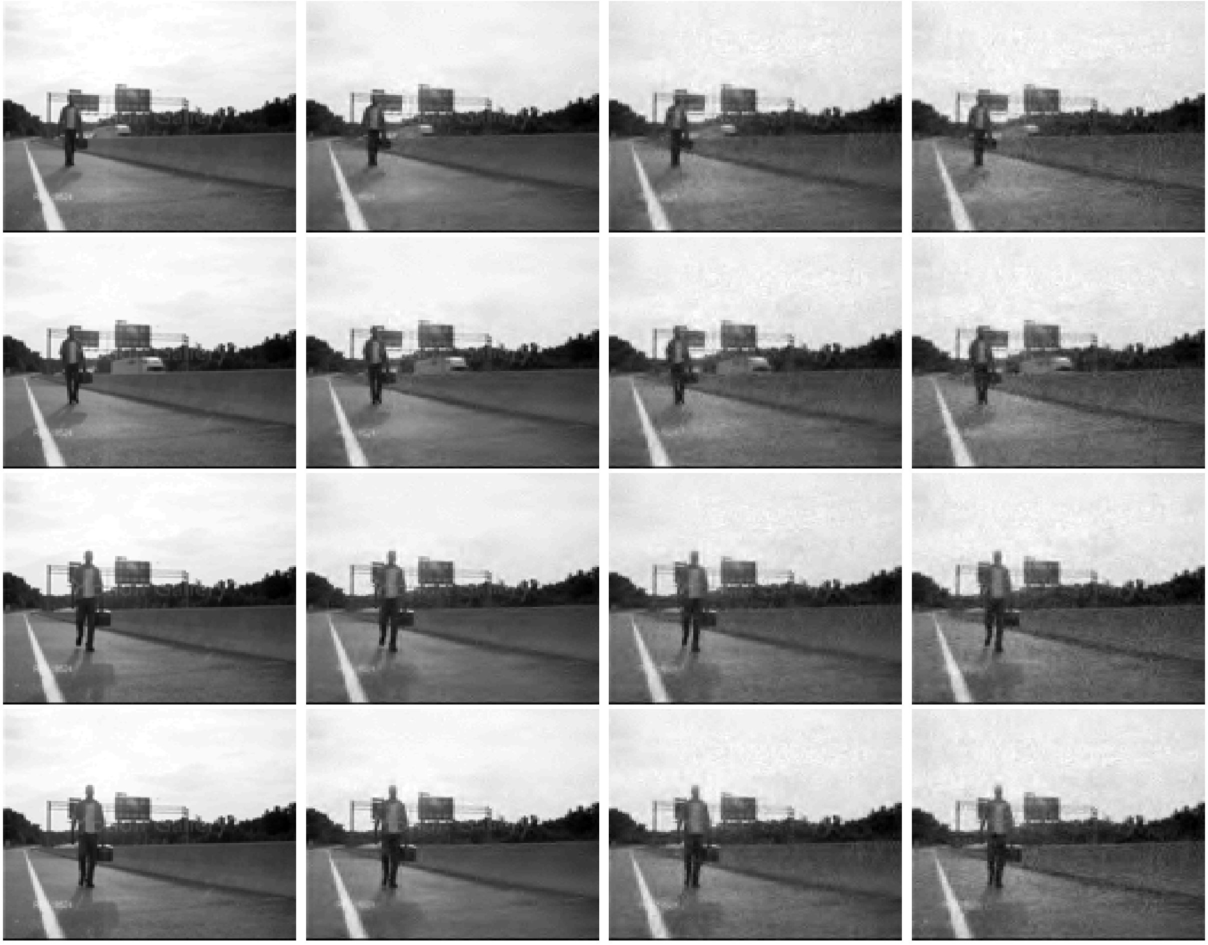
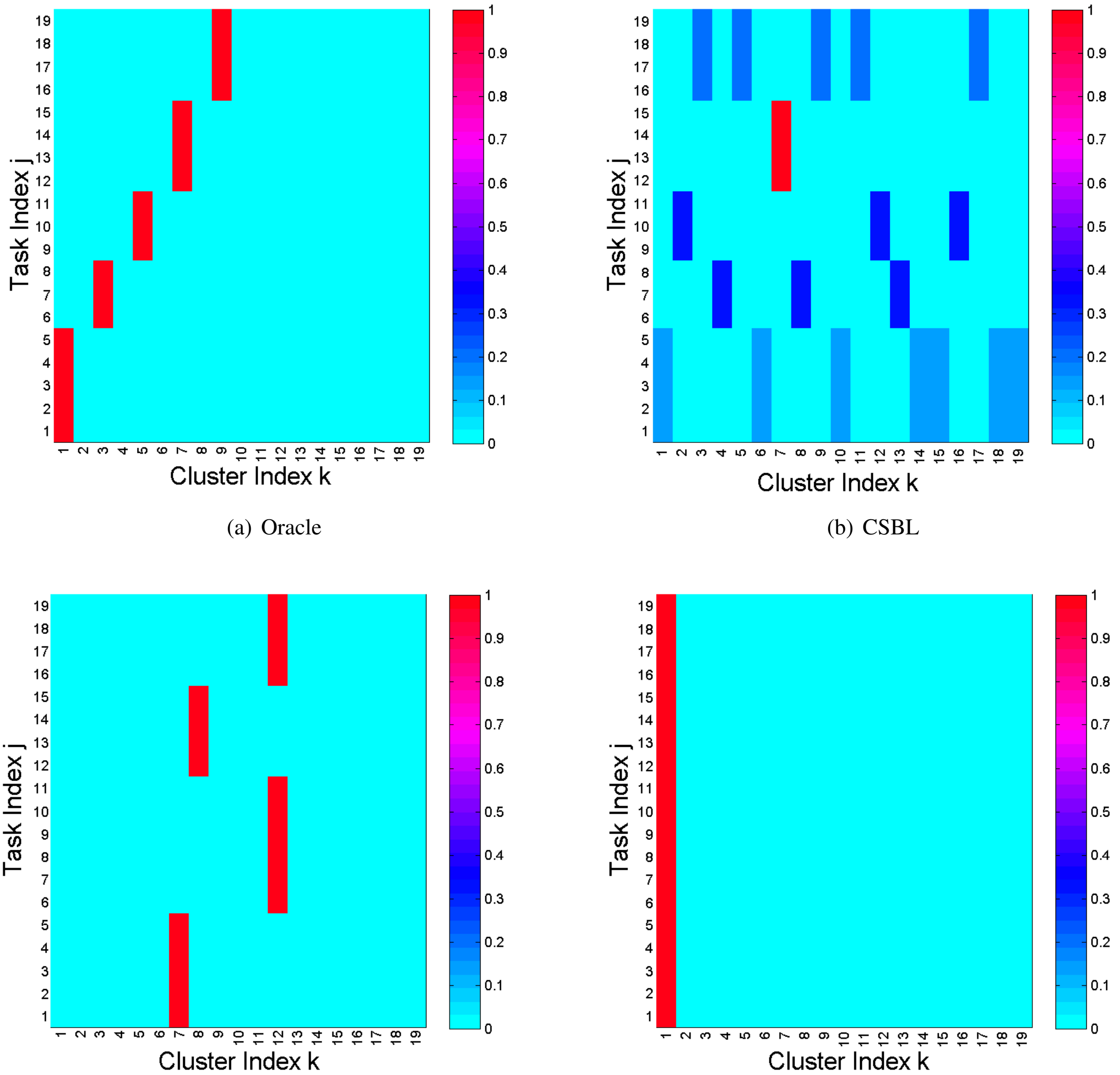


Figure 11: Reconstructions of 128×128 images from dynamic scene 5, cluster size 4. From left to right: Original image, C-SBL, DP, MMV. Sampling rate is $N/M = 0.33$.



(a) Oracle

(b) CSBL

(c) DP

(d) MMV

Figure 12: Heat-map of W matrices for 128×128 image reconstructions. (a) Oracle case (cluster patterns are known), (b) C-SBL, (c) DP, and (d) MMV.

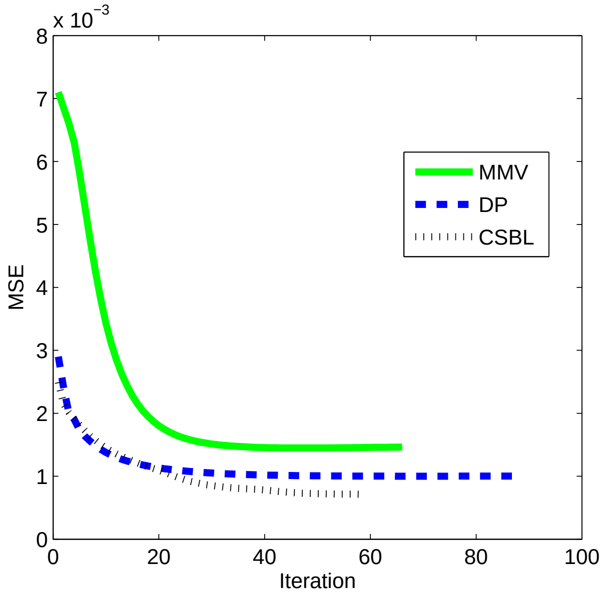


Figure 13: MSE versus iteration for image reconstructions at resolution 128 × 128.

4 Proof of Theorem 2

Here we present the proof of Theorem 2; a few relatively inconsequential technical details are omitted for brevity. Under the stipulated conditions, we are concerned with stationary points of

$$\mathcal{L}_k(\Lambda_k) = \text{tr}\left[Y_{\Omega_k} Y_{\Omega_k}^\top (\Sigma_k)^{-1}\right] + |\Omega_k| \log |\Sigma_k|, \tag{1}$$

where $\Sigma_k = \nu I + \Phi\Lambda_k\Phi^\top$, in the limit when $\nu \to 0$. At any such point, Y_{Ω_k} must be an element of span$[\Phi\Lambda_k^{1/2}]$ or the cost will be driven to infinity. (This occurs for the same reason that $[1/x + \log x] \to \infty$ as $x \to 0$.) In this regard we will first assume that $\Phi\Lambda_k\Phi^\top$ is invertible, which also allows us to simply assume that $\nu = 0$ in (1).

Let $\bar{\Phi}$ be the collection of columns of Φ which correspond with nonzero rows of $X_{\Omega_k}^*$, so it follows that $Y_{\Omega_k} = \bar{\Phi}\bar{X}_{\Omega_k}^*$. Near any candidate stationary point Λ_k, we may express (1) as

$$\mathcal{L}(a,b) = |\Omega_k| \log |a\Sigma_k + b\bar{\Phi}\Delta^2\bar{\Phi}^\top| + \text{tr}\left[Y_{\Omega_k}^\top \left(a\Sigma_k + b\bar{\Phi}\Delta^2\bar{\Phi}^\top\right)^{-1} Y_{\Omega_k}\right], \tag{2}$$

where $\Sigma_k = \Phi\Lambda_k\Phi^\top$ and Δ is an arbitrary positive diagonal matrix. If Λ_k is a stationary point, then it must be that

$$\left.\frac{\partial \mathcal{L}(a,b)}{\partial a}\right|_{a=1,b=0} = 0, \quad \left.\frac{\partial \mathcal{L}(a,b)}{\partial b}\right|_{a=1,b=0} \geq 0, \tag{3}$$

otherwise we could alter a (up or down) or increase b from zero to decrease (1). Let $Z \equiv z(a,b) \triangleq a\Sigma_k + b\bar{\Phi}\Delta^2\bar{\Phi}^\top$. Then

$$\begin{aligned}\frac{\partial \mathcal{L}(a,b)}{\partial a} &= |\Omega_k|\text{tr}\left[Z^{-1}\Sigma_k\right] - \text{tr}\left[Y_{\Omega_k}^\top Z^{-1}\Sigma_k Z^{-1} Y_{\Omega_k}\right] \\ \frac{\partial \mathcal{L}(a,b)}{\partial b} &= |\Omega_k|\text{tr}\left[Z^{-1}\bar{\Phi}\Delta^2\bar{\Phi}^\top\right] - \text{tr}\left[Y_{\Omega_k}^\top Z^{-1}\bar{\Phi}\Delta^2\bar{\Phi}^\top Z^{-1} Y_{\Omega_k}\right].\end{aligned}$$

Since $z(1,0) = \Sigma_k$,

$$\left.\frac{\partial \mathcal{L}(a,b)}{\partial a}\right|_{a=1,b=0} = |\Omega_k|\text{tr}\left[I_N\right] - \text{tr}\left[Y_{\Omega_k}^\top \Sigma_k^{-1} Y_{\Omega_k}\right], \tag{4}$$

$$\left.\frac{\partial \mathcal{L}(a,b)}{\partial b}\right|_{a=1,b=0} = |\Omega_k|\text{tr}\left[\Sigma_k^{-1}\bar{\Phi}\Delta^2\bar{\Phi}^\top\right] - \text{tr}\left[W\Delta^2 W^\top\right], \tag{5}$$

where $W \triangleq Y_{\Omega_k}^\top \Sigma_k^{-1} \bar{\Phi}$. Equating the first gradient equation to zero gives $\mathrm{tr}\left[Y_{\Omega_k}^\top \Sigma_k^{-1} Y_{\Omega_k}\right] = N|\Omega_k|$. For the second we first demonstrate that we may assume $\bar{X}_{\Omega_k}^*$ is invertible without loss of generality. To see this, note that the condition

$$\min_{\Psi > 0} \kappa(\Psi \bar{X}_{\Omega_k}^* (\bar{X}_{\Omega_k}^*)^\top \Psi) < \frac{N}{D} \tag{6}$$

cannot be satisfied unless $\bar{X}_{\Omega_k}^*$ is full row rank, and hence we must have $D \leq |\Omega_k|$. Additionally, if $D < |\Omega_k|$, we can convert (1) to an equivalent objective function with the value of $|\Omega_k|$ reduced to D such that w.l.o.g. $\bar{X}_{\Omega_k}^*$ is now a $D \times D$ full rank matrix. This is possible since (1) only depends on $\bar{\Phi}\bar{X}_{\Omega_k}^*$ through the outer-product $Y_{\Omega_k} Y_{\Omega_k}^\top = \bar{\Phi}\bar{X}_{\Omega_k}^* (\bar{X}_{\Omega_k}^*)^\top \bar{\Phi}^\top$, which can always be reparameterized such that $\bar{X}_{\Omega_k}^*$ is a $D \times D$ full rank matrix.

Consequently, the righthand side of (5) is equivalent to

$$\begin{aligned}
& |\Omega_k| \mathrm{tr}\left[\bar{X}_{\Omega_k}^* (\bar{X}_{\Omega_k}^*)^{-1} \Delta^2 (\bar{X}_{\Omega_k}^*)^{-\top} (\bar{X}_{\Omega_k}^*)^\top \bar{\Phi}^\top \Sigma_k^{-1} \bar{\Phi}\right] - \mathrm{tr}\left[\bar{X}_{\Omega_k}^* (\bar{X}_{\Omega_k}^*)^{-1} \Delta^2 (\bar{X}_{\Omega_k}^*)^{-\top} (\bar{X}_{\Omega_k}^*)^\top W^\top W\right] \\
= \; & |\Omega_k| \mathrm{tr}\left[(\bar{X}_{\Omega_k}^*)^{-1} \Delta^2 (\bar{X}_{\Omega_k}^*)^{-\top} Y_{\Omega_k}^\top \Sigma_k^{-1} Y_{\Omega_k}\right] - \mathrm{tr}\left[(\bar{X}_{\Omega_k}^*)^{-1} \Delta^2 (\bar{X}_{\Omega_k}^*)^{-\top} Y_{\Omega_k}^\top \Sigma_k^{-1} Y_{\Omega_k} Y_{\Omega_k}^\top \Sigma_k^{-1} Y_{\Omega_k}\right] \\
\leq \; & |\Omega_k| \lambda_{max}\left[(\bar{X}_{\Omega_k}^*)^{-1} \Delta^2 (\bar{X}_{\Omega_k}^*)^{-\top}\right] \mathrm{tr}\left[Y_{\Omega_k}^\top \Sigma_k^{-1} Y_{\Omega_k}\right] - \lambda_{min}\left[(\bar{X}_{\Omega_k}^*)^{-1} \Delta^2 (\bar{X}_{\Omega_k}^*)^{-\top}\right] \mathrm{tr}\left[\left(Y_{\Omega_k}^\top \Sigma_k^{-1} Y_{\Omega_k}\right)^2\right] \\
= \; & \frac{|\Omega_k| \mathrm{tr}\left[Y_{\Omega_k}^\top \Sigma_k^{-1} Y_{\Omega_k}\right]}{\lambda_{min}\left[\Delta^{-1} \bar{X}_{\Omega_k}^* (\bar{X}_{\Omega_k}^*)^\top \Delta^{-1}\right]} - \frac{\mathrm{tr}\left[\left(Y_{\Omega_k}^\top \Sigma_k^{-1} Y_{\Omega_k}\right)^2\right]}{\lambda_{max}\left[\Delta^{-1} \bar{X}_{\Omega_k}^* (\bar{X}_{\Omega_k}^*)^\top \Delta^{-1}\right]},
\end{aligned} \tag{7}$$

where the inequality comes from the fact that $\lambda_{min}(A)\mathrm{tr}(B) \leq \mathrm{tr}(AB) \leq \lambda_{max}(A)\mathrm{tr}(B)$ for any positive semi-definite matrices A and B. (Here $\lambda_{min}(A)$ and $\lambda_{max}(A)$ correspond with the smallest and largest eigenvalue of A respectively.)

Now let $\lambda_1, \cdots, \lambda_D$ denote the the eigenvalues of $Y_{\Omega_k}^\top \Sigma_k^{-1} Y_{\Omega_k}$ (where $D = |\Omega_k|$ for the reasons given above), such that $\mathrm{tr}\left[Y_{\Omega_k}^\top \Sigma_k^{-1} Y_{\Omega_k}\right] = \sum_{i=1}^D \lambda_i = N|\Omega_k| = ND$. Also define $A \triangleq \Delta^{-1} \bar{X}_{\Omega_k}^* (\bar{X}_{\Omega_k}^*)^\top \Delta^{-1}$. Then the upper bound from (7) can be modified to

$$D \frac{\sum_{i=1}^D \lambda_i}{\lambda_{min}(A)} - \frac{\sum_{i=1}^D \lambda_i^2}{\lambda_{max}(A)} \leq ND^2 \|A^{-1}\|_2 - \frac{N^2 D}{\|A\|_2}, \tag{8}$$

where the inequality comes from the fact that $\sum_{i=1}^D \lambda_i^2 \geq N^2 D$ given that $\sum_{i=1}^D \lambda_i = ND$.

To summarize then, for a stationary point to occur, it must be that

$$ND^2 \|A^{-1}\|_2 - \frac{N^2 D}{\|A\|_2} \geq \left.\frac{\partial \mathcal{L}(a,b)}{\partial b}\right|_{a=1,b=0} \geq 0. \tag{9}$$

However, if $\kappa(A) < \frac{N}{D}$, then $ND^2\|A^{-1}\|_2 - \frac{N^2 D}{\|A\|_2} < 0$, which means Λ_k cannot be a stationary point. Since Δ can be an arbitrary positive diagonal matrix, we choose $\Delta = \arg\min_\Delta \kappa(A)$ to form the strongest bound. This rules out as a stationary point any Λ_k such that the corresponding Σ_k is full rank. Note that for simplicity we have defined $\Psi \triangleq \Delta^{-1}$ in the original theorem statement as Δ is invertible and the actual parameterization is irrelevant.

We now only need consider the rare $\Lambda_k \neq \Lambda_k^*$ values such that both Y_{Ω_k} is an element of span$[\Phi \Lambda_k^{1/2}]$ and the corresponding Σ_k is *not* full rank. Technically, if Σ_k is not full rank, the cost function (1) is not defined. It is here that careful consideration of the limit of $\nu \to 0$, where the limit is take outside of the minimization, ameliorates the problem. With this in mind, it is straightforward to collapse the problem by projecting Y_{Ω_k} and Φ to a lower-dimensional space such that the resulting Σ_k is now full rank. We may then follow the recipe from above in the resulting lower-dimensional space arriving at a similar conclusion. Hence the only stationary point is a Λ_k^* that is maximally row sparse with $\Lambda_k^* \Phi^\top (\Phi \Lambda_k^* \Phi^\top)^\dagger Y_{\Omega_k} = X_{\Omega_k}^*$.