Appendix A

We consider the linear regression model: given the observed vector $\boldsymbol{y} \in \mathbb{R}^{m \times 1}$, the matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{n}$, where $\boldsymbol{\beta} \in \mathbb{R}^{n \times 1}$ that needs to be determined, and $\boldsymbol{n} \in \mathbb{R}^{m \times 1}$ is the noise vector. We need to find $\boldsymbol{\beta}$ via:

$$\min_{\boldsymbol{\beta}} \; l\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta}\right) \tag{43}$$

where $l(\cdot)$ denotes the loss function, whose expression can be chosen according to the type of noise. Although $l(\cdot)$ may be nonconvex, the solution to (43) should satisfy:

$$\boldsymbol{A}^T l'\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta}\right) = \boldsymbol{0} \tag{44}$$

The iteratively reweighted least squares (IRLS) can be used to solve the regression problem, which is based on viewing $l(\cdot)$ as an adaptively (re-)weighted least squares estimator. For more information about IRLS, please refer to [16]. According to the IRLS, the score function of $l(x)$ can be expressed as:

$$l'(x) = x \cdot w(x) \tag{45}$$

where

$$w(x) = \begin{cases} l'(x)/x, & |x| \neq 0 \\ 1, & |x| = 0 \end{cases} \tag{46}$$

thus, (44) amounts to:

$$\boldsymbol{A}^T \boldsymbol{W}\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta}\right) = \boldsymbol{0} \tag{47}$$

and we obtain:

$$\boldsymbol{\beta} = \left(\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A}\right)^{-1} \boldsymbol{A}^T \boldsymbol{W} \boldsymbol{y} \tag{48}$$

where $\boldsymbol{W} = \operatorname{diag}(w(\boldsymbol{r}_1), \cdots, w(\boldsymbol{r}_m))$ and $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta}$ denotes the residual. Besides, (48) is also the solution to:

$$\min_{\boldsymbol{\beta}} \; \frac{1}{2} \left\| \sqrt{\boldsymbol{W}} \boldsymbol{r} \right\|_2^2 \tag{49}$$

where $\sqrt{\boldsymbol{W}} = \operatorname{diag}(\sqrt{w(\boldsymbol{r}_1)}, \cdots, \sqrt{w(\boldsymbol{r}_m)})$. Therefore, to resist noise, IRLS can be considered to assign different weights to residuals via (49).

It is known that when $\boldsymbol{y}$ is contaminated by Gaussian noise or in the absence of noise, ℓ_2-norm or $l(x) = \frac{x^2}{2}$ is optimal, and the weights for all error are all same and equal to one via (45). While when $\boldsymbol{y}$ is corrupted by big outliers, it is better to assign small weights for big errors. Thus, many nonconvex M-estimators, such as Welsch and Cauchy, are proposed to solve this problem. Those nonconvex M-estimators and their corresponding weights are tabulated in Table VII. The curves of M-estimators are shown in Fig. 10. However, although the nonconvex M-estimators assign small weights to outliers, they also change the weights for the 'normal' data. Here, 'normal' data refer to observations without noise or with only Gaussian noise. To illustrate it, the weight functions of the M-estimators in Table VII are plotted in Fig. 11. When the observations are contaminated by outliers, ℓ_2-norm assigns the same weights for all data including outliers, while although Cauchy and Welsch assign small weights for outliers, they all change the weights for the 'normal' data. Thus, a good M-estimator should keep the weights for 'normal' data fixed, and assigns small weights for outliers. It is seen that the proposed HOW function concurrently satisfies the above needs. For example, it assigns the same weights for x when $|x| \leq 2$, while it assigns small weights for x when $|x| > 2$. The HOW function is the hybrid of ℓ_2-norm and Welsch, because as shown in Fig. 10, Welsch is bounded above and can reject big outliers while Cauchy is unbounded above. That is to say, compared with ℓ_2-norm, although the Cauchy M-estimator is less sensitive to outliers, it introduces infinite energy when the magnitude of noise is infinitely large, because it is not bounded above.

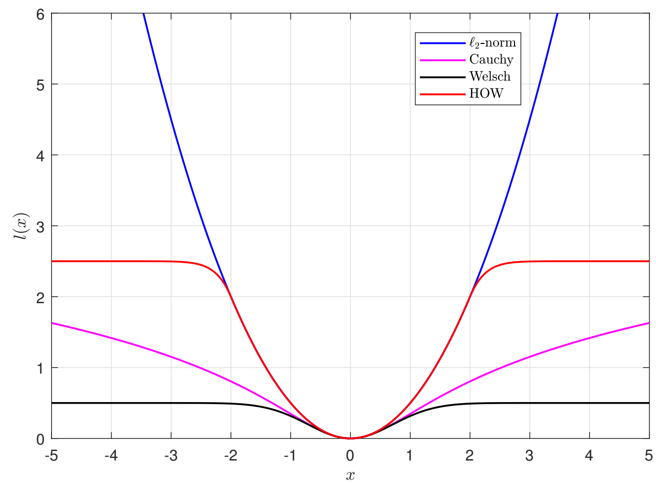

Fig. 10: Different loss M-estimators with $\sigma = 1$ and $c = 2$.

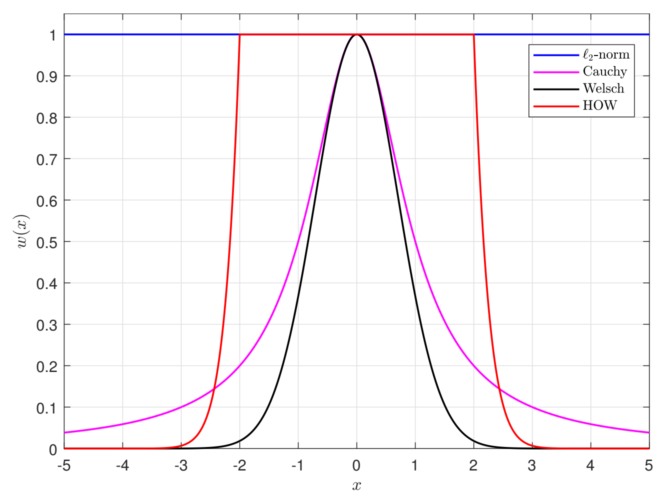

Fig. 11: Different weight functions with $\sigma = 1$ and $c = 2$.

Appendix B

Proof: It follows from [41] that $l_{c,\sigma}(x)$ is an M-estimator because it satisfies:

1) $l_{c,\sigma}(x) \geq 0$;
2) $l_{c,\sigma}(0) = 0$;
3) $l_{c,\sigma}(x) = l_{c,\sigma}(-x)$;
4) $l_{c,\sigma}(x) \geq l_{c,\sigma}(y)$ for $|x| \geq |y|$.

TABLE VII: Different loss functions and their weight functions

	ℓ_2-norm	Cauchy	Welsch	HOW
$l(x)$	$x^2/2$	$\frac{\sigma^2}{2}\log\left(1+(x/\sigma)^2\right)$	$\frac{\sigma^2}{2}\left(1-e^{-\frac{x^2}{\sigma^2}}\right)$	$\begin{cases} x^2/2, & \lvert x\rvert\le c \\ \frac{\sigma^2}{2}\left(1-e^{\frac{c^2-x^2}{\sigma^2}}\right)+\frac{c^2}{2}, & \lvert x\rvert>c \end{cases}$
$w(x)$	1	$\frac{1}{1+(x/\sigma)^2}$	$e^{-\frac{x^2}{\sigma^2}}$	$\begin{cases} 1, & \lvert x\rvert\le c \\ e^{\frac{c^2-x^2}{\sigma^2}}, & \lvert x\rvert>c \end{cases}$

The derivative of $l_{c,\sigma}(x)$ is:

$$\psi(x)=\begin{cases} x, & |x|\le c \\ x\cdot e^{\frac{c^2-x^2}{\sigma^2}}, & |x|>c \end{cases}$$

Since $\lim_{x\to\infty}\psi(x)=0$, that is, $\psi(x)$ is not monotonic, $l_{c,\sigma}(x)$ is a $redescending$ M-estimator [37]. ■

APPENDIX C

Proof: From the expression of $f(x)$, we know that $f(x)$ is a convex function if and only if $f(x)$ with $|x|>c$ is a convex function. Thus, when $|x|>c$,

$$f^{'}(x)=x-x\cdot e^{\frac{c^2-x^2}{\sigma^2}}$$

and

$$\begin{aligned} f^{''}(x)&=1-\left(1-\frac{2x^2}{\sigma^2}\right)e^{\frac{c^2-x^2}{\sigma^2}} \\ &=e^{\frac{c^2-x^2}{\sigma^2}}\left(e^{\frac{x^2-c^2}{\sigma^2}}+\frac{2x^2}{\sigma^2}-1\right) \\ &\ge e^{\frac{c^2-x^2}{\sigma^2}}\left(\frac{x^2-c^2}{\sigma^2}+1+\frac{2x^2}{\sigma^2}-1\right) \\ &=e^{\frac{c^2-x^2}{\sigma^2}}\left(\frac{x^2-c^2}{\sigma^2}+1+\frac{2x^2}{\sigma^2}-1\right) \\ &=e^{\frac{c^2-x^2}{\sigma^2}}\left(\frac{3x^2-c^2}{\sigma^2}\right) \\ &>0 \end{aligned} \tag{50}$$

where the first inequality is obtained using $e^x>x+1$ for any $x\in\mathbb{R}$, and the last inequality is due to the prior condition $x>c$. Therefore, $f(x)$ is a convex function. ■

APPENDIX D

Proof:

1) According to (12), we have:

$$\begin{aligned} \varphi(-y)&=\sup_{x\in\mathbb{R}}\ -\frac{(-y-x)^2}{2}+l_{c,\sigma}(x) \\ &\overset{t=-x}{=}\sup_{t\in\mathbb{R}}\ -\frac{(-y+t)^2}{2}+l_{c,\sigma}(-t) \\ &=\sup_{t\in\mathbb{R}}\ -\frac{(y-t)^2}{2}+l_{c,\sigma}(t) \\ &=\varphi(y) \end{aligned} \tag{51}$$

where the penultimate equation is obtained because $l_{c,\sigma}(x)$ is an even function. Thus, $\varphi(y)$ is symmetric.

2) According to the inversion rule of subgradient relations for Legendre-Ferchel transform and (11), we obtain:

$$\arg\sup_{x}\ x\cdot y-f(x)=\partial f^*(y)=y+\partial\varphi(y) \tag{52}$$

First, we define $q(x)=x\cdot y-f(x)$, which is concave because $f(x)$ is convex, thus the solution $x^\star$ to (52) satisfies $\nabla q(x^\star)=0$, and we have:

$$y=\nabla f(x^\star)=\begin{cases} 0, & |x^\star|\le c \\ x^\star-x^\star\cdot e^{\frac{c^2-(x^\star)^2}{\sigma^2}}, & |x^\star|>c \end{cases} \tag{53}$$

That is to say, when $|y|>0$, $x^\star\in(-\infty,-c)\cup(c,\infty)$, while when $|y|=0$, $x^\star\in[-c,c]$. Besides, we can conclude that the solution $x^\star$ is unique when $|y|>0$ because $f'(x)$ is monotonic for $|x|>c$ due to (50), while it is not unique when $y=0$. Hence $\varphi(y)$ is differentiable except the zero point, and it has a set of subgradients at 0, i.e., $\partial\varphi(0)\in[-c,c]$. Thus, $\varphi(y)$ is nonsmooth at the zero point.
In addition, when $y=0$, we conclude that $0=\varphi(0)+0$ according to (11). Thus, we have $\varphi(0)=0$. For $|y|>0$, i.e., $|x|>c$, as $|x|\to c$, $y\to 0$ and $\varphi(y)\to 0$ as well. Therefore, $\varphi(y)$ is continuous.

3) According to (53), when $y>0$, the solution to $\arg\max_{x}\ y\cdot x-f(x)$ is unique, and it satisfies:

$$y=x^\star-x^\star\cdot e^{\frac{c^2-(x^\star)^2}{\sigma^2}} \tag{54}$$

implying that $y<x^\star$, and by (52) ($x^\star=y+\varphi^{'}(y)$, note that we replace $\partial\varphi(y)$ with $\varphi^{'}(y)$ because $\varphi(y)$ is differentiable for $|y|>0$), we obtain $\varphi^{'}(y)>0$. Therefore, $\varphi(y)$ is increasing for $y>0$ and $\varphi(y)$ is nonnegative because $\varphi(0)=0$.

4) To verify $\varphi(y_1+y_2)\le\varphi(y_1)+\varphi(y_2)$ for any $y_1,y_2\in\mathbb{R}$, we divide it into three cases.
If $y_1\cdot y_2=0$ and we assume that $y_1\neq 0$ and $y_2=0$, it is easy to check $\varphi(y_1+y_2)=\varphi(y_1)\le\varphi(y_1)+\varphi(0)\le\varphi(y_1)+\varphi(y_2)$ due to $\varphi(0)=0$.
If $y_1\cdot y_2>0$ and we first suppose that $y_1>0$ and $y_2>0$. For any $y>0$ (implying $x>c$), we have (54) and by (50), we draw a conclusion that y increases with $x^\star$. Combining $x^\star=y+\varphi^{'}(y)$ and (54), we obtain $\varphi^{'}(y)=x^\star\cdot e^{\frac{c^2-(x^\star)^2}{\sigma^2}}$. We define $r(x)=x\cdot e^{\frac{c^2-x^2}{\sigma^2}}$ and $r^{'}(x)=\left(1-\frac{2x}{\sigma^2}\right)\cdot e^{\frac{c^2-x^2}{\sigma^2}}<0$ because $\frac{2x}{\sigma^2}>1$ when $x>c$, which will be analyzed in Appendix F. Thus, we

know that $\varphi^{'}(y)$ is a decreasing function of y. Then, we obtain:

$$\begin{aligned}&\int_0^{y_2}\varphi^{'}(y_1+y)-\varphi^{'}(y)dy\\=&\varphi(y_1+y_2)-\varphi(y_2)-\varphi(y_1)+\varphi(0)\\\leq&0\end{aligned}$$

Thus, $\varphi(y_1+y_2)\leq\varphi(y_1)+\varphi(y_2)$. On the other hand, if $y_1\cdot y_2>0$, $y_1<0$ and $y_2<0$, we employ (51) to yield:

$$\begin{aligned}\varphi(y_1+y_2)=&\varphi(-y_1-y_2)\\&\leq\varphi(-y_1)+\varphi(-y_2)\\&=\varphi(y_1)+\varphi(y_2)\end{aligned}$$

If $y_1\cdot y_2<0$ and assuming that $y_1>0$ and $y_2<0$, we can draw the conclusion immediately from $\varphi(y_1+y_2)\leq\varphi(y_1+|y_2|)\leq\varphi(y_1)+\varphi(|y_2|)=\varphi(y_1)+\varphi(y_2)$. ∎

Appendix E
Proof of Proposition 1

Proof: Since $f^*(y)$ is convex in (11), we know that $yx-f^*(y)=yx-\varphi(y)-\frac{y^2}{2}$ and $-g(y)=-\frac{(y-x)^2}{2}-\varphi(y)$ in (13) w.r.t. y are concave, thus $g(y)$ is convex. In addition, the solution to (14) is equal to that to (13), and we have by the inversion rule for subgradient relations:

$$\partial f(x)=\arg\sup_y\ y\cdot x-f^*(y)$$

thus,

$$y=\nabla f(x)=\begin{cases}0, & |x|\leq c\\ x-x\cdot e^{(c^2-x^2)/\sigma^2}, & |x|>c\end{cases}$$

Moreover, it is easy to check that $\nabla f(x)$ is an odd function and increases with x when $|x|>c$ according to (50), thus we conclude that $\nabla f(x)$ is monotonic, that is, $p_\varphi(x)$ is non-decreasing. ∎

Appendix F
Proof of Proposition 2

Proof: For convenience, only $x\geq0$ is considered. We define $\Delta b=y-p_\varphi(y)$, which is used to measure the bias. As shown in Fig. 12, ℓ_0-'norm' as the regularizer is unbiased while ℓ_1-norm has a constant bias for $y>1$, that is to say, $\Delta b=0$ for the ℓ_0-'norm' and Δb is a constant for the ℓ_1-norm. To bridge the gap between them, it is necessary to ensure that Δb decreases with y ($y>c$, $c=1$ in Fig. 12). Thus, we have by Proposition 1 :

$$\begin{aligned}\Delta b&=y-\left(y-y\cdot e^{\frac{c^2-y^2}{\sigma^2}}\right)\\&=y\cdot e^{\frac{c^2-y^2}{\sigma^2}}\end{aligned}\tag{55}$$

and we need:

$$\frac{d(\Delta b)}{dy}=\left(1-\frac{2y^2}{\sigma^2}\right)\cdot e^{\frac{c^2-y^2}{\sigma^2}}<0\tag{56}$$

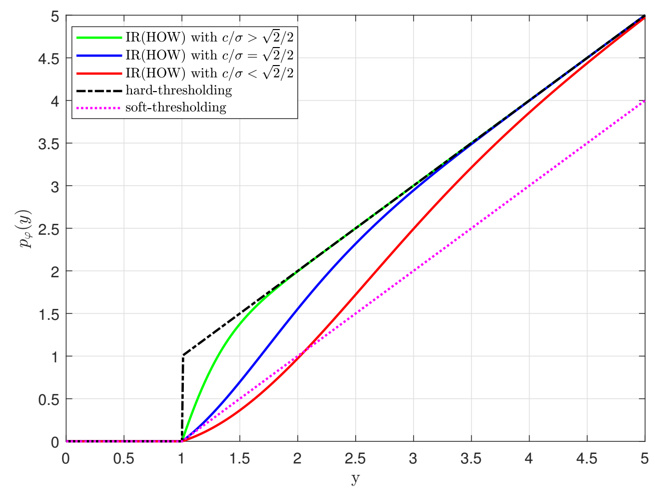

Fig. 12: Comparison of $p_\varphi(y)$ for different regularizers.

Therefore, for any $y>c$, we obtain:

$$1-\left(2y^2\right)/\sigma^2<0\tag{57a}$$

$$y>\frac{\sigma}{\sqrt{2}}\tag{57b}$$

Since $y>c$, we get $c\geq\frac{\sigma}{\sqrt{2}}$. This conclusion makes sure that $\varphi(y)$ satisfies the triangle inequality in Proposition D. It can be seen in Fig. 12 that when $c<\frac{\sigma}{\sqrt{2}}$, Δb increases first and then decreases with y, while when $c\geq\frac{\sigma}{\sqrt{2}}$, Δb decreases with y. This completes the proof. ∎

Appendix G
Proof of Theorem 1

Proof: There are five parameters, i.e., c, e, $\boldsymbol{S}$, $\boldsymbol{U}$ and $\boldsymbol{V}$, in Algorithm 1, which will be theoretically analyzed that updating each of them in the proposed algorithm leads to the decrease of objective function.

First, when only updating c from c_1 to c_2 ($c_2<c_1$) and fixing the remaining parameters, we divide $x\in\mathbb{R}$ into $|x_1|\leq c_2$, $c_2<|x_2|\leq c_1$ and $|x_3|>c_1$, then we have

$$\begin{aligned}&l_{c_1,\sigma}(x)-l_{c_2,\sigma}(x)\\=&\ l_{c_1,\sigma}(x_1)+l_{c_1,\sigma}(x_2)+l_{c_1,\sigma}(x_3)\\&-l_{c_2,\sigma}(x_1)-l_{c_2,\sigma}(x_2)-l_{c_2,\sigma}(x_3)\\=&\ a_1+a_2+a_3\end{aligned}$$

where $a_1=l_{c_1,\sigma}(x_1)-l_{c_2,\sigma}(x_1)$, $a_2=l_{c_1,\sigma}(x_2)-l_{c_2,\sigma}(x_2)$ and $a_3=l_{c_1,\sigma}(x_3)-l_{c_2,\sigma}(x_3)$. According to the definition of

$l_{c,\sigma}(x)$ in (7), we get:

$$\begin{aligned} &l_{c_1,\sigma}(x_1) = l_{c_2,\sigma}(x_1) = x_1^2/2 \\ &l_{c_1,\sigma}(x_2) = x_2^2/2 \\ &l_{c_1,\sigma}(x_3) = \frac{\sigma^2}{2}\left(1 - e^{\frac{c_1^2 - x_3^2}{\sigma^2}}\right) + \frac{c_1^2}{2} \\ &l_{c_2,\sigma}(x_2) = \frac{\sigma^2}{2}\left(1 - e^{\frac{c_2^2 - x_2^2}{\sigma^2}}\right) + \frac{c_2^2}{2} \\ &l_{c_2,\sigma}(x_3) = \frac{\sigma^2}{2}\left(1 - e^{\frac{c_2^2 - x_3^2}{\sigma^2}}\right) + \frac{c_2^2}{2} \end{aligned} \tag{59a}$$

Thus, $a_1 = 0$. Besides, According to Appendix C, for any $c_2 > 0$, a_2 increases with x_2, thus we have

$$a_2 > l_{c_1,\sigma}(c_2) - l_{c_2,\sigma}(c_2) = 0 \tag{60}$$

When $|x_3| > c_1$, $\frac{\partial l_{c,\sigma}(x)}{\partial c} = c \cdot \left(1 - e^{\left(c^2 - x^2\right)/\sigma^2}\right) > 0$, thus $l_{c_1,\sigma}(x_3) > l_{c_2,\sigma}(x_3)$ due to $c_1 > c_2$, resulting in $a_3 > 0$. Therefore, we get:

$$l_{c_1,\sigma}(x) - l_{c_2,\sigma}(x) > 0$$

Hence, $l_{c,\sigma}(x)$ increases with c.

Second, when just updating σ, since σ is not in the range, we take the derivative of $l_{c,\sigma}(x)$ w.r.t. σ. If $|x| \leq c$, $\frac{\partial l_{c,\sigma}(x)}{\partial \sigma} = 0$, while if $|x| > c$, $\frac{\partial l_{c,\sigma}(x)}{\partial \sigma} = \sigma + \left(\frac{c^2 - x^2}{\sigma} - \sigma\right) e^{\frac{c^2 - x^2}{\sigma^2}} = \sigma\left(1 - \left(\frac{x^2 - c^2}{\sigma^2} + 1\right) e^{\frac{c^2 - x^2}{\sigma^2}}\right) \overset{d}{\geq} \sigma\left(1 - e^{\frac{x^2 - c^2}{\sigma^2}} e^{\frac{c^2 - x^2}{\sigma^2}}\right) = 0$, where d holds because $e^y \geq y + 1$ for any $y \in \mathbb{R}$, thus we obtain $\frac{\partial l_{c,\sigma}(x)}{\partial \sigma} \geq 0$.

Therefore, for the non-increasing c^{k+1} and σ^{k+1}, i.e., $c^{k+1} \leq c^k$ and $\sigma^{k+1} \leq \sigma^k$, $l_{c^{k+1},\sigma^{k+1}}(x) \leq l_{c^k,\sigma^k}(x)$ holds. Note that the impact of c and e on (16) and (18) is the same. Thus, we have:

$$\mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^k\right) \leq \mathcal{C}_{c^k,\sigma^k}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^k\right) \tag{61}$$

For fixed c^{k+1} and σ^{k+1}, based on the properties of SCAD method and (20), we have:

$$\begin{aligned} &\mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^{k+1}, \boldsymbol{V}^{k+1}, \boldsymbol{S}^{k+1}\right) \\ &\leq \mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^{k+1}, \boldsymbol{V}^k, \boldsymbol{S}^{k+1}\right) \\ &\leq \mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^{k+1}\right) \\ &\leq \mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^k\right) \end{aligned} \tag{62}$$

Combining (61) and (62), for the non-increasing c and σ, which is satisfied owing to (34), the sequence $\left\{\mathcal{C}_{c^k,\sigma^k}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^k\right),\ k = 1, 2, \cdots\right\}$ converges as $k \to \infty$ since $\mathcal{C}_{c^k,\sigma^k}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^k\right)$ is bounded below. This completes the proof. ■

Appendix H
Proof of Theorem 2

Proof: We first state the definition of critical point.

Definition 1. *If $0 \in \partial f(x)$, then x is a critical point of f [45].*

Then, according to Theorem 1, we know that:

$$\mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^k\right) \leq \mathcal{C}_{c^k,\sigma^k}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^k\right) \tag{63}$$

Since (14) is a convex problem, and the solution in Proposition 1 satisfies:

$$0 \in \partial g(y^\star) \tag{64}$$

Thus, when updating $\boldsymbol{S}^{k+1}$ with the use of $\boldsymbol{U}^k$ and $\boldsymbol{V}^k$, we have:

$$\mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^{k+1}\right) \leq \mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^k\right) \tag{65}$$

$$\mathbf{0} \in \partial_{\boldsymbol{S}} \mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^{k+1}\right) \tag{66}$$

where $\partial_{\boldsymbol{S}} \mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^{k+1}\right) = \frac{\partial \mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}\right)}{\partial \boldsymbol{S}}|_{\boldsymbol{S}=\boldsymbol{S}^{k+1}}$.

Besides, SASD is used to solve $\boldsymbol{U}$ and $\boldsymbol{V}$, and the decrease of $h(\boldsymbol{U}, \boldsymbol{V}^k)$ along $\boldsymbol{U}$ can be obtained exactly:

$$\begin{aligned} &h(\boldsymbol{U}^{k+1}, \boldsymbol{V}^k) \\ =&\frac{1}{2}\left\|\boldsymbol{H}_\Omega^{k+1} - \left(\left(\boldsymbol{U}^k - \widetilde{\mu}_{\boldsymbol{U}}^k \widetilde{\nabla} h_{\boldsymbol{V}^k}(\boldsymbol{U}^k)\right)\boldsymbol{V}^k\right)_\Omega\right\|_F^2 \\ =&\frac{1}{2}\left\|\boldsymbol{H}_\Omega^{k+1} - \left(\boldsymbol{U}^k\boldsymbol{V}^k\right)_\Omega + \left(\widetilde{\mu}_{\boldsymbol{U}}^k \widetilde{\nabla} h_{\boldsymbol{V}^k}(\boldsymbol{U}^k)\boldsymbol{V}^k\right)_\Omega\right\|_F^2 \\ =&\frac{1}{2}\left\|\boldsymbol{H}_\Omega^{k+1} - \left(\boldsymbol{U}^k\boldsymbol{V}^k\right)_\Omega\right\|_F^2 + \frac{1}{2}\left\|\left(\widetilde{\mu}_{\boldsymbol{U}}^k \widetilde{\nabla} h_{\boldsymbol{V}^k}(\boldsymbol{U}^k)\boldsymbol{V}^k\right)_\Omega\right\|_F^2 \\ &+ \left\langle \boldsymbol{H}_\Omega^{k+1} - \left(\boldsymbol{U}^k\boldsymbol{V}^k\right)_\Omega, \left(\widetilde{\mu}_{\boldsymbol{U}}^k \widetilde{\nabla} h_{\boldsymbol{V}^k}(\boldsymbol{U}^k)\boldsymbol{V}^k\right)_\Omega\right\rangle \\ \overset{a}{=}&\frac{1}{2}\left\|\boldsymbol{H}_\Omega^{k+1} - \left(\boldsymbol{U}^k\boldsymbol{V}^k\right)_\Omega\right\|_F^2 - \frac{1}{2}\left(\widetilde{\mu}_{\boldsymbol{U}}^k\right)^2\left\|\left(\widetilde{\nabla} h_{\boldsymbol{V}^k}(\boldsymbol{U}^k)\boldsymbol{V}^k\right)_\Omega\right\|_F^2 \\ =&\frac{1}{2}\left\|\boldsymbol{H}_\Omega^{k+1} - \left(\boldsymbol{U}^k\boldsymbol{V}^k\right)_\Omega\right\|_F^2 - \frac{\left|\left\langle \nabla h_{\boldsymbol{V}^k}(\boldsymbol{U}^k), \widetilde{\nabla} h_{\boldsymbol{V}^k}(\boldsymbol{U}^k)\right\rangle\right|^2}{2\left\|\left(\widetilde{\nabla} h_{\boldsymbol{V}^k}(\boldsymbol{U}^k)\boldsymbol{V}^k\right)_\Omega\right\|_F^2} \end{aligned} \tag{67}$$

where a is due to the optimal step size $\widetilde{\mu}_{\boldsymbol{U}}^k$. Thus, by adding $\varphi(\boldsymbol{S}_\Omega^{k+1})$ on the both sides of (67), we have:

$$\begin{aligned} \mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^{k+1}, \boldsymbol{V}^k, \boldsymbol{S}^{k+1}\right) =&\mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^k, \boldsymbol{V}^k, \boldsymbol{S}^{k+1}\right) \\ &- \frac{\left|\left\langle \nabla h_{\boldsymbol{V}^k}(\boldsymbol{U}^k), \widetilde{\nabla} h_{\boldsymbol{V}^k}(\boldsymbol{U}^k)\right\rangle\right|^2}{2\left\|\left(\widetilde{\nabla} h_{\boldsymbol{V}^k}(\boldsymbol{U}^k)\boldsymbol{V}^k\right)_\Omega\right\|_F^2} \end{aligned} \tag{68}$$

Similarly, the decrease of $h(\boldsymbol{U}^{k+1}, \boldsymbol{V})$ along $\boldsymbol{V}$ can be computed exactly:

$$\begin{aligned} &h(\boldsymbol{U}^{k+1}, \boldsymbol{V}^{k+1}) \\ =&\frac{1}{2}\left\|\boldsymbol{H}_\Omega^{k+1} - \left(\boldsymbol{U}^{k+1}\boldsymbol{V}^k\right)_\Omega\right\|_F^2 \\ &- \frac{\left|\left\langle \nabla h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k), \widetilde{\nabla} h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k)\right\rangle\right|^2}{2\left\|\left(\boldsymbol{U}^{k+1}\widetilde{\nabla} h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k)\right)_\Omega\right\|_F^2} \end{aligned} \tag{69}$$

and we have:

$$\mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^{k+1},\boldsymbol{V}^{k+1},\boldsymbol{S}^{k+1}\right)=\mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^{k+1},\boldsymbol{V}^{k},\boldsymbol{S}^{k+1}\right) - \frac{\left|\left\langle \nabla h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k), \widetilde{\nabla} h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k)\right\rangle\right|^2}{2\left\|\left(\boldsymbol{U}^{k+1}\widetilde{\nabla} h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k)\right)_\Omega\right\|_F^2} \tag{70}$$

Adding (63), (65), (68) and (70), we obtain:

$$\begin{aligned}&\mathcal{C}_{c^{k},\sigma^{k}}\left(\boldsymbol{U}^{k},\boldsymbol{V}^{k},\boldsymbol{S}^{k}\right)-\mathcal{C}_{c^{k+1},\sigma^{k+1}}\left(\boldsymbol{U}^{k+1},\boldsymbol{V}^{k+1},\boldsymbol{S}^{k+1}\right)\\ \geq&\frac{\left|\left\langle \nabla h_{\boldsymbol{V}^{k}}(\boldsymbol{U}^k), \widetilde{\nabla} h_{\boldsymbol{V}^{k}}(\boldsymbol{U}^k)\right\rangle\right|^2}{2\left\|\left(\widetilde{\nabla} h_{\boldsymbol{V}^{k}}(\boldsymbol{U}^k)\boldsymbol{V}^k\right)_\Omega\right\|_F^2}\\ &+\frac{\left|\left\langle \nabla h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k), \widetilde{\nabla} h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k)\right\rangle\right|^2}{2\left\|\left(\boldsymbol{U}^{k+1}\widetilde{\nabla} h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k)\right)_\Omega\right\|_F^2}\end{aligned} \tag{71}$$

Summing (71) from $k=0$ to $N-1$, where N is a positive integer:

$$\begin{aligned}&\mathcal{C}_{c^{0},\sigma^{0}}\left(\boldsymbol{U}^{0},\boldsymbol{V}^{0},\boldsymbol{S}^{0}\right)-\mathcal{C}_{c^{N},\sigma^{N}}\left(\boldsymbol{U}^{N},\boldsymbol{V}^{N},\boldsymbol{S}^{N}\right)\\ \geq&\sum_{k=0}^{N-1}\frac{\left|\left\langle \nabla h_{\boldsymbol{V}^{k}}(\boldsymbol{U}^k), \widetilde{\nabla} h_{\boldsymbol{V}^{k}}(\boldsymbol{U}^k)\right\rangle\right|^2}{2\left\|\left(\widetilde{\nabla} h_{\boldsymbol{V}^{k}}(\boldsymbol{U}^k)\boldsymbol{V}^k\right)_\Omega\right\|_F^2}\\ &+\sum_{k=0}^{N-1}\frac{\left|\left\langle \nabla h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k), \widetilde{\nabla} h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k)\right\rangle\right|^2}{2\left\|\left(\boldsymbol{U}^{k+1}\widetilde{\nabla} h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k)\right)_\Omega\right\|_F^2}\end{aligned} \tag{72}$$

Since $\mathcal{C}_{c,\sigma}\left(\boldsymbol{U},\boldsymbol{V},\boldsymbol{S}\right)$ is nonnegative and bounded by 0, we have:

$$\lim_{k\to\infty}\frac{\left|\left\langle \nabla h_{\boldsymbol{V}^{k}}(\boldsymbol{U}^k), \widetilde{\nabla} h_{\boldsymbol{V}^{k}}(\boldsymbol{U}^k)\right\rangle\right|}{\left\|\left(\widetilde{\nabla} h_{\boldsymbol{V}^{k}}(\boldsymbol{U}^k)\boldsymbol{V}^k\right)_\Omega\right\|_F}=0 \tag{73}$$

and

$$\lim_{k\to\infty}\frac{\left|\left\langle \nabla h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k), \widetilde{\nabla} h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k)\right\rangle\right|}{\left\|\left(\boldsymbol{U}^{k+1}\widetilde{\nabla} h_{\boldsymbol{U}^{k+1}}(\boldsymbol{V}^k)\right)_\Omega\right\|_F}=0 \tag{74}$$

In addition, assuming that $\|\boldsymbol{U}^0\|_F<\infty$, $\|\boldsymbol{V}^0\|_F<\infty$ and $\|\boldsymbol{S}^0\|_F<\infty$, and suppose that $\left\{\boldsymbol{U}^k,\boldsymbol{V}^k\right\}$ are of full rank, namely, $(\boldsymbol{U}^T\boldsymbol{U})^{-1}$ and $(\boldsymbol{V}\boldsymbol{V}^T)^{-1}$ exist, we obtain that $\widetilde{\nabla} h_{\boldsymbol{V}}(\boldsymbol{U})$ and $\widetilde{\nabla} h_{\boldsymbol{U}}(\boldsymbol{V})$ are bounded since they can be calculated by matrix addition, multiplication and inverse operators. Moreover, the step sizes $\widetilde{\mu}_{\boldsymbol{U}}$ and $\widetilde{\mu}_{\boldsymbol{V}}$ are bounded, thus $\{\boldsymbol{U}^k,\boldsymbol{V}^k\}$ are bounded because of (32) and (33), and $\|\boldsymbol{S}^k\|_F<\infty$ since the solution to $\boldsymbol{S}^k$ depends on $\{\boldsymbol{U}^k,\boldsymbol{V}^k\}$ by (20).

Let $\left\{\left(\boldsymbol{U}^{k_j},\ \boldsymbol{V}^{k_j},\ \boldsymbol{S}^{k_j}\right)\right\}$ be a subsequence of $\left\{\left(\boldsymbol{U}^{k},\boldsymbol{V}^{k},\boldsymbol{S}^{k}\right)\right\}$ such that $\lim_{k_j\to\infty}\boldsymbol{U}^{k_j}=\boldsymbol{U}^\star$, $\lim_{k_j\to\infty}\boldsymbol{V}^{k_j}=\boldsymbol{V}^\star$ and $\lim_{k_j\to\infty}\boldsymbol{S}^{k_j}=\boldsymbol{S}^\star$. After obtaining (73) and (74), according to Lemma 4.2 and Lemma 4.3 in [44], we get:

$$\lim_{k_j\to\infty}\nabla h_{\boldsymbol{V}^{k_j}}(\boldsymbol{U}^{k_j})=\boldsymbol{0} \tag{75}$$

and

$$\lim_{k_j\to\infty}\nabla h_{\boldsymbol{U}^{k_j}}(\boldsymbol{V}^{k_j})=\boldsymbol{0} \tag{76}$$

Combining (75), (76) and (66), we have:

$$\begin{aligned}\boldsymbol{0}&=\lim_{k_j\to\infty}\nabla h_{\boldsymbol{V}^{k_j}}(\boldsymbol{U}^{k_j})\\ &=\lim_{k_j\to\infty}\partial_{\boldsymbol{U}}\mathcal{C}_{c^{k_j},\sigma^{k_j}}\left(\boldsymbol{U}^{k_j},\boldsymbol{V}^{k_j},\boldsymbol{S}^{k_j}\right)\end{aligned} \tag{77}$$

$$\begin{aligned}\boldsymbol{0}&=\lim_{k_j\to\infty}\nabla h_{\boldsymbol{U}^{k_j}}(\boldsymbol{V}^{k_j})\\ &=\lim_{k_j\to\infty}\partial_{\boldsymbol{V}}\mathcal{C}_{c^{k_j},\sigma^{k_j}}\left(\boldsymbol{U}^{k_j},\boldsymbol{V}^{k_j},\boldsymbol{S}^{k_j}\right)\end{aligned} \tag{78}$$

$$\begin{aligned}\boldsymbol{0}&\in\lim_{k_j\to\infty}\partial_{\boldsymbol{S}}\mathcal{C}_{c^{k_j},\sigma^{k_j}}\left(\boldsymbol{U}^{k_j-1},\boldsymbol{V}^{k_j-1},\boldsymbol{S}^{k_j}\right)\\ &=\lim_{k_j\to\infty}\partial_{\boldsymbol{S}}\mathcal{C}_{c^{k_j},\sigma^{k_j}}\left(\boldsymbol{U}^{k_j},\boldsymbol{V}^{k_j},\boldsymbol{S}^{k_j}\right)\end{aligned} \tag{79}$$

where $\partial_{\boldsymbol{U}}\mathcal{C}_{c,\sigma}\left(\boldsymbol{U},\boldsymbol{V},\boldsymbol{S}\right)=\frac{\partial\mathcal{C}_{c,\sigma}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{S})}{\partial\boldsymbol{U}}$ and $\partial_{\boldsymbol{V}}\mathcal{C}_{c,\sigma}\left(\boldsymbol{U},\boldsymbol{V},\boldsymbol{S}\right)=\frac{\partial\mathcal{C}_{c,\sigma}(\boldsymbol{U},\boldsymbol{V},\boldsymbol{S})}{\partial\boldsymbol{V}}$. Equation (79) is obtained because $\lim_{k_j\to\infty}\boldsymbol{U}^{k_j}-\boldsymbol{U}^{k_j-1}=-\widetilde{\mu}_{\boldsymbol{U}}^{k_j-1}\widetilde{\nabla} h_{\boldsymbol{V}^{k_j-1}}(\boldsymbol{U}^{k_j-1})=\boldsymbol{0}$ and $\lim_{k_j\to\infty}\boldsymbol{V}^{k_j}-\boldsymbol{V}^{k_j-1}=-\widetilde{\mu}_{\boldsymbol{V}}^{k_j-1}\widetilde{\nabla} h_{\boldsymbol{U}^{k_j}}(\boldsymbol{V}^{k_j-1})=\boldsymbol{0}$ according to (32) and (33), respectively. Therefore, the limit point of the subsequence $\left\{\left(\boldsymbol{U}^{k_j},\boldsymbol{V}^{k_j},\boldsymbol{S}^{k_j}\right)\right\}$, namely, $(\boldsymbol{U}^\star,\boldsymbol{V}^\star,\boldsymbol{S}^\star)$, is a critical point of (18). ■

APPENDIX I
SHADOW AND SPECULARITY REMOVAL FROM FACES

Fig. 13: Shadow and specularity removal from faces. Face-5 to Face-8 from $Subject$ 01 are original images, corrupted images by 10 dB Gaussian noise, 10 dB GMM noise and occlusions, respectively. Similarly, Face-9 to Face-12 from $Subject$ 20 are original images, corrupted images by 10 dB Gaussian noise, 10 dB GMM noise and occlusions, respectively. Images from the second row to the last are recovered images by RPCA-HOW, RPCA-Welsch, IALM, SSGoDec and NCRPCA, respectively.

TABLE VIII: Runtime comparison for different algorithms.

Method	Face-5		Face-6		Face-7		Face-8		Face-9		Face-10		Face-11		Face-12	
	Iter.	Time	Iter.	Time	Iter.	Time	Iter.	Time	Iter.	Time	Iter.	Time	Iter.	Time	Iter.	Time
RPCA-HOW	**11.25**	**1.274**	**7.85**	**0.828**	**10.3**	**1.135**	**12.6**	**1.460**	**12.55**	**1.326**	**8.75**	**0.862**	**10.55**	**1.171**	**16.35**	**1.950**
RPCA-Welsch	27.75	3.378	13.75	1.507	19.65	2.271	26.6	3.162	24	2.787	13.5	1.488	18.95	2.140	26.7	3.106
IALM	239	18.361	174.85	12.911	198.30	14.365	218.9	16.921	258	19.439	187.85	14.121	205.20	15.445	234.70	18.352
SSGoDec	101	2.629	101	2.634	101	2.649	101	2.652	101	2.653	101	2.645	101	2.663	101	2.662
NCRPCA	58	4.879	96.9	8.313	100	2.649	57	4.739	58	4.917	99.7	8.704	100	8.875	57	4.807

APPENDIX J
BACKGROUND MODELING FROM SURVEILLANCE VIDEO

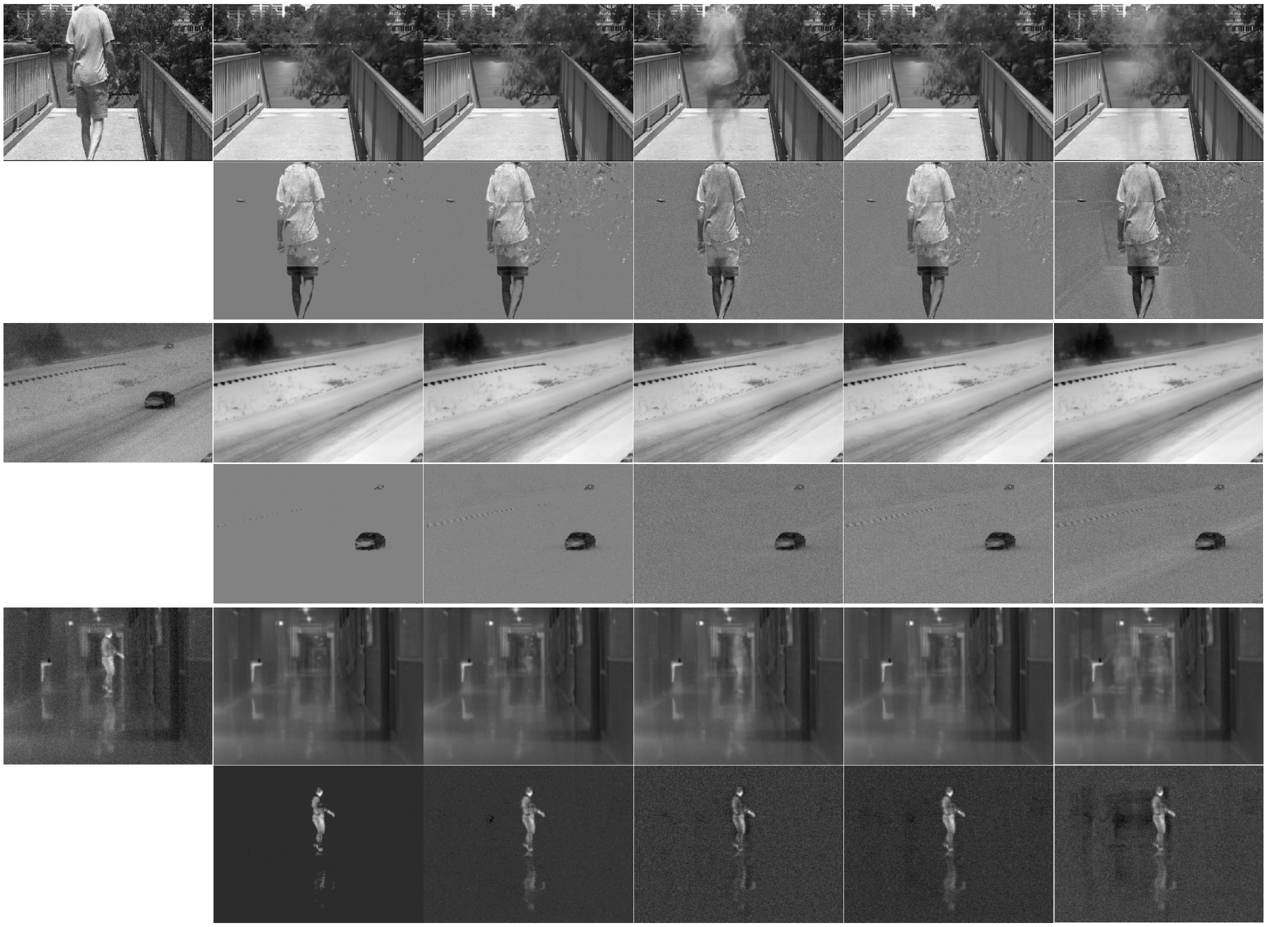

Fig. 14: Background and foreground separation results of different algorithms using real videos. Images from left to right are original images, recovered results by RPCA-HOW, RPCA-Welsch, IALM, SSGoDec and NCRPCA, respectively.

TABLE IX: Quantitative results on background and foreground separation

Method	Overpass				Blizzard				Corridor			
	Iter.	Time	$\frac{\|\boldsymbol{S}\|_0}{mn}$	$\mathrm{F_m}$	Iter.	Time	$\frac{\|\boldsymbol{S}\|_0}{mn}$	$\mathrm{F_m}$	Iter.	Time	$\frac{\|\boldsymbol{S}\|_0}{mn}$	$\mathrm{F_m}$
RPCA-HOW	**9**	**6.24**	**0.089**	**0.5712**	**7**	**4.86**	**0.029**	**0.7208**	**9**	**6.19**	**0.033**	**0.6005**
RPCA-Welsch	11	7.88	1	0.5683	9	6.91	1	0.7198	12	8.68	1	0.5915
IALM	130	97.3	0.916	0.4154	116	91.0	0.960	0.7098	161	116	0.960	0.5176
SSGoDec	101	18.9	0.931	0.5664	101	20.4	0.925	0.7174	101	19.0	0.914	0.5931
NCRPCA	100	87.4	1	0.4848	65	58.0	1	0.6632	100	87.1	1	0.5726