Supplementary Material to "Robust Unsupervised Video Foreground Segmentation via Context-Aware Bayesian Tensor Factorization"

Appendix A: Proof for Proposition 1

Proposition 1. *Problem*

$$\max_{q(\Theta)} \mathcal{Q} = \mathcal{L}(q(\Theta)) - \mathcal{R}(q(\Theta)) \tag{A.1}$$

has a close form solution for each iteration of VB inference algorithm. Let $\Theta = \{\Theta_1, \Theta_2, \cdots, \Theta_J\}$ *denote all the variables, and* $\mathcal{R}(q(\Theta_j)) = -\mathbb{E}[h(\Theta_j)]$ *denotes the regularization term for variable* Θ_j*. The logarithm of the distribution* $q_j(\Theta_j)$ *is:*

$$ln(q_j(\Theta_j)) = \mathbb{E}_{\Theta_{i|i\neq j}} ln\big(p(\Theta, \boldsymbol{\mathcal{Y}})\big) + h(\Theta_j) + Constant. \tag{A.2}$$

Proof. The first term in (A.1) is a lower bound of the model evidence $\ln p(\boldsymbol{\mathcal{Y}})$, and the lower bound can be re-written as:

$$\mathcal{L}(q(\Theta)) = \int q(\Theta) \ln\big(\frac{p(\Theta, \boldsymbol{\mathcal{Y}})}{q(\Theta)}\big) d\Theta. \tag{A.3}$$

And the second term in (A.1) is a regularization term for variables in Θ, which can be expressed in a general form:

$$\rho\mathcal{R}(q(\Theta)) = -\mathbb{E}[h(\Theta)] = -\int q(\Theta) h(\Theta) d\theta, \tag{A.4}$$

Thus, (A.1) can be expressed as:

$$\begin{aligned}
&\mathcal{Q} = \mathcal{L}(q(\Theta)) - \rho\mathcal{R}(q(\Theta)) \\
&= \int q(\Theta)\ln\big(\frac{p(\Theta, \boldsymbol{\mathcal{Y}})}{q(\Theta)}\big)d\Theta + \int q(\Theta)h(\Theta)d\theta \\
&= \int q(\Theta)\ln\big(p(\Theta, \boldsymbol{\mathcal{Y}})\big)d\Theta - \int q(\Theta)\ln\big(q(\Theta)\big)d\Theta + \int q(\Theta)h(\Theta)d\theta.
\end{aligned}$$

By applying mean field approximation [1]:

$$q(\Theta) = \prod_{i=1}^{J} q_i(\Theta_i),$$

the objective function $\mathcal{Q}$ becomes:

$$\begin{aligned}
&\mathcal{Q} = \int_{\Theta_1}\int_{\Theta_2}\cdots\int_{\Theta_J}\prod_{i=1}^{J} q_i(\Theta_i)\ln\big(p(\Theta, \boldsymbol{\mathcal{Y}})\big)d\Theta_1 d\Theta_2\cdots d\Theta_J \\
&-\int_{\Theta_1}\int_{\Theta_2}\cdots\int_{\Theta_J}\prod_{i=1}^{J} q_i(\Theta_i)\sum_{i=1}^{J}\ln(q_i(\Theta_i))d\Theta_1 d\Theta_2\cdots d\Theta_J \\
&+\int_{\Theta_1}\int_{\Theta_2}\cdots\int_{\Theta_J}\prod_{i=1}^{J} q_i(\Theta_i)(h(\Theta_j))d\Theta_1 d\Theta_2\cdots d\Theta_J.
\end{aligned}$$

Assume regularization is applied on the variable Θ_j, terms related to Θ_j can be isolated:

$$\begin{aligned}
&\mathcal{Q} = \int_{\Theta_j} q_j(\Theta_j)\int_{\Theta_{i|i\neq j}}\prod_{i=1,i\neq j}^{J} q_i(\Theta_i)\ln\big(p(\Theta, \boldsymbol{\mathcal{Y}})\big)d\Theta_1 d\Theta_2\cdots d\Theta_J \\
&-\int_{\Theta_1}\int_{\Theta_2}\cdots\int_{\Theta_J}\prod_{i=1}^{J} q_i(\Theta_i)\ln(q_j(\Theta_j))d\Theta_1 d\Theta_2\cdots d\Theta_J \\
&-\int_{\Theta_1}\int_{\Theta_2}\cdots\int_{\Theta_J}\prod_{i=1}^{J} q_i(\Theta_i)\sum_{k=1,k\neq j}^{J}\ln(q_k(\Theta_k))d\Theta_1 d\Theta_2\cdots d\Theta_J \\
&+\int_{\Theta_j} q_j(\Theta_j)(h(\Theta_j))\int_{\Theta_{i|i\neq j}}\prod_{i=1,i\neq j}^{J} q_i(\Theta_i)d\Theta_1 d\Theta_2\cdots d\Theta_J
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\Theta_j} q_j(\Theta_j)\mathbb{E}_{\Theta_{i|i\neq j}}\ln\big(p(\Theta,\boldsymbol{\mathcal{Y}})\big)d\Theta_j \\
&- \int_{\Theta_j} q_j(\Theta_j)\ln(q_j(\Theta_j))d\Theta_j \int_{\Theta_{i|i\neq j}} \prod_{i=1,i\neq j}^{J} q_i(\Theta_i)\,\underbrace{d\Theta_1 d\Theta_2\cdots d\Theta_J}_{\text{without } d\Theta_j} \\
&- \int_{\Theta_j} q_j(\Theta_j)d\Theta_j \int_{\Theta_{i|i\neq j}} \prod_{i=1,i\neq j}^{J} q_i(\Theta_i) \sum_{k=1,k\neq j}^{J} \ln(q_k(\Theta_k))\,\underbrace{d\Theta_1 d\Theta_2\cdots d\Theta_J}_{\text{without } d\Theta_j} \\
&+ \int_{\Theta_j} q_j(\Theta_j)(h(\Theta_j))d\Theta_j \int_{\Theta_{i|i\neq j}} \prod_{i=1,i\neq j}^{J} q_i(\Theta_i)\,\underbrace{d\Theta_1 d\Theta_2\cdots d\Theta_J}_{\text{without } d\Theta_j}.
\end{aligned}
$$

By the property of probability density function (PDF), $\int_{\Theta_i} q_i(\Theta_i)d\Theta_i = 1$. The objective function $\mathcal{Q}$ can be simplified as:

$$
\begin{aligned}
&\mathcal{Q} = \int_{\Theta_j} q_j(\Theta_j)\mathbb{E}_{\Theta_{i|i\neq j}}\ln\big(p(\Theta,\boldsymbol{\mathcal{Y}})\big)d\Theta_j \\
&- \int_{\Theta_j} q_j(\Theta_j)\ln(q_j(\Theta_j))d\Theta_j \\
&- \int_{\Theta_{i|i\neq j}} \prod_{i=1,i\neq j}^{J} q_i(\Theta_i) \sum_{k=1,k\neq j}^{J} \ln(q_k(\Theta_k))\,\underbrace{d\Theta_1 d\Theta_2\cdots d\Theta_J}_{\text{without } d\Theta_j} \\
&+ \int_{\Theta_j} q_j(\Theta_j)(h(\Theta_j))d\Theta_j \\
&\underbrace{- \sum_{i=1}^{J} \lambda_i(\int q_i(\theta_i)d\theta_i - 1)}_{\text{PDF property constraint}}.
\end{aligned}
$$

The maximization of the objective function $\mathcal{Q}$ respect to $q_j(\Theta_j)$ can be achieved by taking the derivative of $\mathcal{Q}$ and set it to 0:

$$
\begin{aligned}
\frac{\partial\mathcal{Q}}{\partial q_j(\Theta_j)} &= \mathbb{E}_{\Theta_{i|i\neq j}}\ln\big(p(\Theta,\boldsymbol{\mathcal{Y}})\big) - \ln(q_j(\Theta_j)) - 0 + (h(\Theta_j)) + \text{Constant} \\
&= 0.
\end{aligned}
$$

After rearranging the above, the logarithm of the distribution $q_j(\Theta_j)$ is:

$$
\ln(q_j(\Theta_j)) = \mathbb{E}_{\Theta_{i|i\neq j}}\ln\big(p(\Theta,\boldsymbol{\mathcal{Y}})\big) + (h(\Theta_j)) + \text{Constant}, \tag{A.5}
$$

which is the (A.2), so the proof is done.

□

Appendix B: Derivations of Updated Posterior Distributions in Section II-C

In this appendix, the derivation of updating parameters in (8)-(11) are shown. The purpose of updating parameters in (8)-(11) is to update the approximated posteriors and make them approximate the true posteriors.

According to Proposition 1, the logarithm of updated posteriors in each iteration can be obtained by (A.2). Specifically, the logarithm of $q_j(\Theta_j)$ equals the sum of the expectation of logarithm of joint distribution respect to all the parameters expect the variable Θ_j, the regularization on the $q_j(\Theta_j)$, and a constant. The first term in (A.2) can be obtained as the following:

The joint distribution is the product of likelihood function (3) and priors mentioned in Section II-B.2. Specifically, the joint distribution has the following form:

$$p(\Theta|\boldsymbol{\mathcal{Y}}) = p\big(\boldsymbol{\mathcal{Y}}|\{\mathbf{A}^{(n)}\}_{n=1}^{N}, \boldsymbol{\mathcal{F}}, \tau\big)\prod_{n=1}^{N} p(\mathbf{A}^{(n)}|\boldsymbol{\lambda})p(\boldsymbol{\mathcal{F}}|\boldsymbol{\gamma})p(\boldsymbol{\lambda})p(\boldsymbol{\gamma})p(\tau), \tag{A.6}$$

where the exact form of the likelihood function (3) is

$$p\big(\boldsymbol{\mathcal{Y}}|\{\mathbf{A}^{(n)}\}_{n=1}^{N}, \boldsymbol{\mathcal{F}}, \tau\big) = \prod_{i_1=1}^{I_1}\cdots\prod_{i_N=1}^{I_N}\mathcal{N}\big(\boldsymbol{\mathcal{Y}}_{i_1\dots i_N}|\langle \mathbf{a}_{i_1}^{(1)}, \dots, \mathbf{a}_{i_N}^{(N)}\rangle + \boldsymbol{\mathcal{F}}_{i_1\dots i_N}, \tau^{-1}\big)$$

$$= \prod_{i_1=1}^{I_1}\cdots\prod_{i_N=1}^{I_N}\Big((2\pi\tau^{-1})^{-\frac{1}{2}}e^{-\frac{1}{2}(\boldsymbol{\mathcal{Y}}_{i_1\dots i_N}-\left\langle \mathbf{a}_{i_1}^{(1)},\dots,\mathbf{a}_{i_N}^{(N)}\right\rangle-\boldsymbol{\mathcal{F}}_{i_1\dots i_N})\tau(\boldsymbol{\mathcal{Y}}_{i_1\dots i_N}-\left\langle \mathbf{a}_{i_1}^{(1)},\dots,\mathbf{a}_{i_N}^{(N)}\right\rangle-\boldsymbol{\mathcal{F}}_{i_1\dots i_N})}\Big); \tag{A.7}$$

and the priors are used can be found in [2], where the priors of factor matrices $\{\mathbf{A}^{(n)}\}_{n=1}^{N}$ are

$$p(\mathbf{A}^{(n)}|\boldsymbol{\lambda}) = \prod_{i_n=1}^{I_n}\mathcal{N}\big(\mathbf{a}_{i_n}^{(n)}|\mathbf{0}, \boldsymbol{\Lambda}^{-1}\big) = \prod_{i_n=1}^{I_n}\big(\det(2\pi\boldsymbol{\Lambda}^{-1})^{\frac{1}{2}}e^{-\frac{1}{2}\mathbf{a}_{i_n}^{(n)T}\boldsymbol{\Lambda}\mathbf{a}_{i_n}^{(n)}}\big), \qquad n = 1, \cdots, N. \tag{A.8}$$

With the assumption of shared precision for all dimensions, the prior of precision of factor matrices is

$$p(\boldsymbol{\lambda}) = \prod_{r=1}^{R}\mathrm{Ga}(\lambda_r|c_0, d_0) = \prod_{r=1}^{R}\frac{d_0^{c_0}}{\Gamma(c_0)}\lambda_r^{c_0-1}e^{-d_0\lambda_r}. \tag{A.9}$$

The prior of foreground $\boldsymbol{\mathcal{F}}$ is

$$p(\boldsymbol{\mathcal{F}}|\boldsymbol{\gamma}) = \prod_{i_1,\dots,i_N}\mathcal{N}\big(\boldsymbol{\mathcal{F}}_{i_1\dots i_N}|0, \gamma_{i_1\dots i_N}^{-1}\big) = \prod_{i_1,\dots,i_N}\big((2\pi\gamma_{i_1\dots i_N}^{-1})^{-\frac{1}{2}}e^{\frac{1}{2}\boldsymbol{\mathcal{F}}_{i_1\dots i_N}\gamma_{i_1\dots i_N}\boldsymbol{\mathcal{F}}_{i_1\dots i_N}}\big); \tag{A.10}$$

and the prior of precision of foreground is

$$p(\boldsymbol{\gamma}) = \prod_{i_1,\dots,i_N}\mathrm{Ga}\big(\gamma_{i_1\dots i_N}|a_0^{\gamma}, b_0^{\gamma}\big) = \prod_{i_1,\dots,i_N}\frac{(b_0^{\gamma})^{a_0^{\gamma}}}{\Gamma(a_0^{\gamma})}\gamma_{i_1\dots i_N}^{a_0^{\gamma}-1}e^{-b_0^{\gamma}\gamma_{i_1\dots i_N}}. \tag{A.11}$$

Lastly, the prior of noise precision is

$$p(\tau) = \mathrm{Ga}\big(\tau|a_0^\tau, b_0^\tau\big) = \frac{(b_0^\tau)^{a_0^\tau}}{\Gamma(a_0^\tau)}\tau^{a_0^\tau-1}e^{-b_0^\tau\tau}. \tag{A.12}$$

According to (A.6)-(A.12), the logarithm of the joint distribution $\ln\big(p(\Theta,\boldsymbol{\mathcal{Y}})\big)$ can be expressed as:

$$\begin{aligned}
\ln\big(p(\Theta,\boldsymbol{\mathcal{Y}})\big) &= \sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}\big(\frac{1}{2}\ln(\tau) - \frac{\tau}{2}(\boldsymbol{\mathcal{Y}}_{i_1\ldots i_N} - \langle \mathbf{a}_{i_1}^{(1)},\ldots,\mathbf{a}_{i_N}^{(N)}\rangle - \boldsymbol{\mathcal{F}}_{i_1\ldots i_N})^2\big)\\
&+\sum_{i_n=1}^{I_n}\big(\frac{1}{2}\ln(|\boldsymbol{\Lambda}|) - \frac{1}{2}\mathbf{a}_{i_n}^{(n)T}\boldsymbol{\Lambda}\mathbf{a}_{i_n}^{(n)}\big)\\
&+\sum_{i_1,\ldots,i_N}\big(\frac{1}{2}\ln(\gamma_{i_1\ldots i_N}) - \frac{1}{2}\gamma_{i_1\ldots i_N}(\boldsymbol{\mathcal{F}}_{i_1\ldots i_N})^2\big) + \sum_{r=1}^{R}\big((c_0-1)\ln(\lambda_r) - d_0\lambda_r\big)\\
&+\sum_{i_1,\ldots,i_N}\big((a_0^\gamma-1)\ln(\gamma_{i_1\ldots i_N}) - b_0^\gamma\gamma_{i_1\ldots i_N}\big) + (a_0^\tau-1)\ln(\tau) - b_0^\tau\tau + \text{Constant}.
\end{aligned} \tag{A.13}$$

By the close form solution of logarithm of updated posteriors (A.2) and the logarithm of joint distribution (A.13), the updated posterior distribution of each variables can be found.

The updated posterior distributions of factor matrices $\{\mathbf{A}^{(n)}\}_{n=1}^N$ are

$$\begin{aligned}
\ln(q(\mathbf{a}_{i_n}^{(n)})) &= \mathbb{E}_{\Theta\backslash\mathbf{a}_{i_n}^{(n)}}\ln\big(p(\Theta,\boldsymbol{\mathcal{Y}})\big) + \text{Constant}\\
&= \mathbb{E}\left[-\frac{\tau}{2}(\boldsymbol{\mathcal{Y}}_{i_1\ldots i_N} - \langle \mathbf{a}_{i_1}^{(1)},\ldots,\mathbf{a}_{i_N}^{(N)}\rangle - \boldsymbol{\mathcal{F}}_{i_1\ldots i_N})^2 - \frac{1}{2}\mathbf{a}_{i_n}^{(n)T}\boldsymbol{\Lambda}\mathbf{a}_{i_n}^{(n)}\right] + \text{Constant}\\
&= -\frac{1}{2}\big(\mathbf{a}_{i_n}^{(n)T}(\mathbb{E}_q[\tau]\mathbb{E}_q\left[\mathbf{A}_{i_n}^{(\backslash n)T}\mathbf{A}_{i_n}^{(\backslash n)}\right] + \mathbb{E}_q[\boldsymbol{\Lambda}])\mathbf{a}_{i_n}^{(n)}\big)\\
&+ \mathbf{a}_{i_n}^{(n)T}\mathbb{E}_q[\tau]\mathbb{E}_q\left[\mathbf{A}_{i_n}^{(\backslash n)T}\right]\mathrm{vec}\big(\boldsymbol{\mathcal{Y}} - \mathbb{E}_q[\boldsymbol{\mathcal{F}}]\big) + \text{Constant}, \qquad n = 1,\cdots,N.
\end{aligned}$$

From the above logarithm probability density, $q(\mathbf{a}_{i_n}^{(n)})$ follows a Gaussian distribution and the posterior parameters are:

$$\begin{aligned}
\hat{\mathbf{a}}_{i_n}^{(n)} &= \mathbb{E}_q[\tau]\mathbf{V}_{i_n}^{(n)}\mathbb{E}_q\left[\mathbf{A}_{i_n}^{(\backslash n)T}\right]\mathrm{vec}\big(\boldsymbol{\mathcal{Y}} - \mathbb{E}_q[\boldsymbol{\mathcal{F}}]\big),\\
\mathbf{V}_{i_n}^{(n)} &= \big(\mathbb{E}_q[\tau]\mathbb{E}_q\left[\mathbf{A}_{i_n}^{(\backslash n)T}\mathbf{A}_{i_n}^{(\backslash n)}\right] + \mathbb{E}_q[\boldsymbol{\Lambda}]\big)^{-1}, \qquad n = 1,\cdots,N,
\end{aligned} \tag{A.14}$$

which are the same as (8).

The updated posterior distributions of foreground $\boldsymbol{\mathcal{F}}$ is

$$\begin{aligned}
\ln(q(\boldsymbol{\mathcal{F}})) &= \mathbb{E}_{\Theta\setminus\boldsymbol{\mathcal{F}}}\ln\big(p(\Theta,\boldsymbol{\mathcal{Y}})\big) + (h(\boldsymbol{\mathcal{F}})) + \text{Constant}\\
&= \mathbb{E}\Big[\sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}\big(\frac{1}{2}\ln(\tau) - \frac{\tau}{2}(\boldsymbol{\mathcal{Y}}_{i_1\ldots i_N} - \big\langle \mathbf{a}_{i_1}^{(1)},\ldots,\mathbf{a}_{i_N}^{(N)}\big\rangle - \boldsymbol{\mathcal{F}}_{i_1\ldots i_N})^2\big)\\
&- \sum_{i_1,\ldots,i_N}\big(\frac{1}{2}\gamma_{i_1\ldots i_N}(\boldsymbol{\mathcal{F}}_{i_1\ldots i_N})^2\big)\Big] - \rho\|\mathrm{vec}(\boldsymbol{\mathcal{F}})\|_1 + \text{Constant}\\
&= \sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}\Big[-\frac{1}{2}(\mathbb{E}[\gamma_{i_1\ldots i_N}] + \mathbb{E}[\tau])(\boldsymbol{\mathcal{F}}_{i_1\ldots i_N})^2 + \mathbb{E}[\tau]\big(\boldsymbol{\mathcal{Y}}_{i_1\ldots i_N} - \mathbb{E}[\big\langle \mathbf{a}_{i_1}^{(1)},\ldots,\mathbf{a}_{i_N}^{(N)}\big\rangle]\big)\boldsymbol{\mathcal{F}}_{i_1\ldots i_N}\Big]\\
&- \rho_1\sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}|\boldsymbol{\mathcal{F}}_{i_1\ldots i_N}| - \rho_2\sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}(n\boldsymbol{\mathcal{F}}^2_{i_1\ldots i_N} - 2(\sum_{*\in\Omega_{i_1\ldots i_N}}\boldsymbol{\mathcal{F}}_*)\boldsymbol{\mathcal{F}}_{i_1\ldots i_N}) + \text{Constant}.
\end{aligned}$$

where n is the total number of entries in the neighborhood $\Omega_{i_1\ldots i_N}$. As $\boldsymbol{\mathcal{F}}_{i_1\ldots i_N}$ can be either positive or negative, the above posterior distribution has two cases:

$$\ln(q(\boldsymbol{\mathcal{F}})) = \begin{cases}
\text{Constant} - \dfrac{1}{2}\displaystyle\sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}(\mathbb{E}[\gamma_{i_1\ldots i_N}] + \mathbb{E}[\tau] + 2n\rho_2)(\boldsymbol{\mathcal{F}}_{i_1\ldots i_N})^2\\
+\displaystyle\sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}\big(\mathbb{E}[\tau]\boldsymbol{\mathcal{Y}}_{i_1\ldots i_N} - \mathbb{E}[\tau]\mathbb{E}[\big\langle \mathbf{a}_{i_1}^{(1)},\ldots,\mathbf{a}_{i_N}^{(N)}\big\rangle] + 2\rho_2(\sum_{*\in\Omega_{i_1\ldots i_N}}\boldsymbol{\mathcal{F}}_*) - \rho_1\big)\boldsymbol{\mathcal{F}}_{i_1\ldots i_N},\\
\text{if } \boldsymbol{\mathcal{F}}_{i_1\ldots i_N} \geq 0;\\
\\
\text{Constant} - \dfrac{1}{2}\displaystyle\sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}(\mathbb{E}[\gamma_{i_1\ldots i_N}] + \mathbb{E}[\tau] + 2n\rho_2)(\boldsymbol{\mathcal{F}}_{i_1\ldots i_N})^2\\
+\displaystyle\sum_{i_1=1}^{I_1}\cdots\sum_{i_N=1}^{I_N}\big(\mathbb{E}[\tau]\boldsymbol{\mathcal{Y}}_{i_1\ldots i_N} - \mathbb{E}[\tau]\mathbb{E}[\big\langle \mathbf{a}_{i_1}^{(1)},\ldots,\mathbf{a}_{i_N}^{(N)}\big\rangle] + 2\rho_2(\sum_{*\in\Omega_{i_1\ldots i_N}}\boldsymbol{\mathcal{F}}_*) + \rho_1\big)\boldsymbol{\mathcal{F}}_{i_1\ldots i_N},\\
\text{if } \boldsymbol{\mathcal{F}}_{i_1\ldots i_N} < 0.
\end{cases}$$

For each of the case, $q(\boldsymbol{\mathcal{F}}_{i_1\ldots i_N})$ follows a Gaussian distribution and the posterior parameters are:

$$\hat{\sigma}^2_{i_1\ldots i_N} = (\mathbb{E}_q[\gamma_{i_1\ldots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2)^{-1}, \tag{A.15}$$

$$\hat{\boldsymbol{\mathcal{F}}}_{i_1\ldots i_N} = \begin{cases}
\hat{\sigma}^2_{i_1\ldots i_N}\big(\mathbb{E}_q[\tau]\boldsymbol{\mathcal{Y}}_{i_1\ldots i_N} - \mathbb{E}_q[\tau]\mathbb{E}_q\big[\big\langle \mathbf{a}_{i_1}^{(1)},\ldots,\mathbf{a}_{i_N}^{(N)}\big\rangle\big] + 2\rho_2(\displaystyle\sum_{*\in\Omega_{i_1\ldots i_N}}\boldsymbol{\mathcal{F}}_*) - \rho_1\big), & \text{if } \boldsymbol{\mathcal{F}}_{i_1\ldots i_N} \geq 0;\\
\hat{\sigma}^2_{i_1\ldots i_N}\big(\mathbb{E}_q[\tau]\boldsymbol{\mathcal{Y}}_{i_1\ldots i_N} - \mathbb{E}_q[\tau]\mathbb{E}_q\big[\big\langle \mathbf{a}_{i_1}^{(1)},\ldots,\mathbf{a}_{i_N}^{(N)}\big\rangle\big] + 2\rho_2(\displaystyle\sum_{*\in\Omega_{i_1\ldots i_N}}\boldsymbol{\mathcal{F}}_*) + \rho_1\big), & \text{if } \boldsymbol{\mathcal{F}}_{i_1\ldots i_N} < 0.
\end{cases}$$

And the above can be re-write as:

$$\hat{\boldsymbol{\mathcal{F}}}_{i_1\dots i_N} = S_{\rho_1}\big(\tilde{\boldsymbol{\mathcal{F}}}_{i_1\dots i_N}\big), \tag{A.16}$$

where

$$\tilde{\boldsymbol{\mathcal{F}}}_{i_1\dots i_N} = \hat{\sigma}^2_{i_1\dots i_N}\big(\mathbb{E}_q[\tau]\boldsymbol{\mathcal{Y}}_{i_1\dots i_N} - \mathbb{E}_q[\tau]\mathbb{E}_q\big[\big\langle \mathbf{a}^{(1)}_{i_1},\dots,\mathbf{a}^{(N)}_{i_N}\big\rangle\big] + 2\rho_2(\sum_{*\in\Omega_{i_1\dots i_N}} \boldsymbol{\mathcal{F}}_*)\big), \tag{A.17}$$

$S(.)$ denotes a shrinkage operator which is defined as: $S_{\rho>0}(x) = \mathrm{sgn}(x)max(|x| - \rho, 0)$.

Therefore, (A.15)-(A.17) are the same as (9)-(11). For other hyperparameters $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$, and τ, their updated posterior parameters can be derived similarly. The detailed derivations are omitted and can be found in [2], as they are not the emphasis of this paper.

Appendix C: Proof for Proposition 2

Proposition 2. *Let $\boldsymbol{\mathcal{F}}^O_{i_1\dots i_N}$ denote the ordinary estimator of the foreground variable $\boldsymbol{\mathcal{F}}_{i_1\dots i_N}$ without any regularization ($\rho_1 = \rho_2 = 0$). Let $\bar{\boldsymbol{\mathcal{F}}}_{\Omega_{i_1\dots i_N}}$ denote the average of foreground estimators for all the neighbors in $\Omega_{i_1\dots i_N}$. $\tilde{\boldsymbol{\mathcal{F}}}_{i_1\dots i_N}$ is a weighted average:*

$$\tilde{\boldsymbol{\mathcal{F}}}_{i_1\dots i_N} = (1 - \frac{2n\rho_2}{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2})\boldsymbol{\mathcal{F}}^O_{i_1\dots i_N} + \frac{2n\rho_2}{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2}\bar{\boldsymbol{\mathcal{F}}}_{\Omega_{i_1\dots i_N}}, \tag{A.18}$$

Proof. From (A.15)-(A.17), the ordinary estimators are obtained when $\rho_1 = \rho_2 = 0$:

$$\boldsymbol{\mathcal{F}}^O_{i_1\dots i_N} = \frac{1}{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau]}\big(\mathbb{E}_q[\tau]\boldsymbol{\mathcal{Y}}_{i_1\dots i_N} - \mathbb{E}_q[\tau]\mathbb{E}_q\big[\big\langle \mathbf{a}^{(1)}_{i_1},\dots,\mathbf{a}^{(N)}_{i_N}\big\rangle\big], \tag{A.19}$$

$$(\sigma^O_{i_1\dots i_N})^2 = (\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau])^{-1}. \tag{A.20}$$

Then (A.17) can be re-written as:

$$\begin{aligned}
\tilde{\boldsymbol{\mathcal{F}}}_{i_1\dots i_N} =& \hat{\sigma}^2_{i_1\dots i_N}\big(\mathbb{E}_q[\tau]\boldsymbol{\mathcal{Y}}_{i_1\dots i_N} - \mathbb{E}_q[\tau]\mathbb{E}_q\big[\big\langle \mathbf{a}^{(1)}_{i_1},\dots,\mathbf{a}^{(N)}_{i_N}\big\rangle\big] + 2\rho_2(\sum_{*\in\Omega_{i_1\dots i_N}} \boldsymbol{\mathcal{F}}_*)\big)\\
=& \frac{1}{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2}\big((\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau])\boldsymbol{\mathcal{F}}^O_{i_1\dots i_N} + 2\rho_2(\sum_{*\in\Omega_{i_1\dots i_N}} \boldsymbol{\mathcal{F}}_*)\big)\\
=& \frac{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau]}{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2}\boldsymbol{\mathcal{F}}^O_{i_1\dots i_N} + \frac{2\rho_2}{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2}(\sum_{*\in\Omega_{i_1\dots i_N}} \boldsymbol{\mathcal{F}}_*)\\
=& \frac{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau]}{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2}\boldsymbol{\mathcal{F}}^O_{i_1\dots i_N} + \frac{2n\rho_2}{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2}\bar{\boldsymbol{\mathcal{F}}}_{\Omega_{i_1\dots i_N}}\\
=& (1 - \frac{2n\rho_2}{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2})\boldsymbol{\mathcal{F}}^O_{i_1\dots i_N} + \frac{2n\rho_2}{\mathbb{E}_q[\gamma_{i_1\dots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2}\bar{\boldsymbol{\mathcal{F}}}_{\Omega_{i_1\dots i_N}}.
\end{aligned}$$

The above is same as (A.18), so the proof is done.

□

Appendix D: Proof for Proposition 3

Proposition 3. *Let $\sigma^O_{i_1\ldots i_N}$ denote the ordinary estimator of standard deviation of foreground variable $\boldsymbol{\mathcal{F}}_{i_1\ldots i_N}$ without any regularization ($\rho_1 = \rho_2 = 0$). To average the $\boldsymbol{\mathcal{F}}_{i_1\ldots i_N}$ with all its neighbors evenly, ρ_2 can be determined by:*

$$\rho_2 = \frac{1}{2(\sigma^O_{i_1\ldots i_N})^2}. \tag{A.21}$$

Proof. It is a fact that $\boldsymbol{\mathcal{F}}_{i_1\ldots i_N}$ is averaged with all its neighbors evenly if

$$\tilde{\boldsymbol{\mathcal{F}}}_{i_1\ldots i_N} = \frac{\boldsymbol{\mathcal{F}}^O_{i_1\ldots i_N} + \sum_{*\in\Omega_{i_1\ldots i_N}} \boldsymbol{\mathcal{F}}_*}{n+1} = \frac{1}{n+1}\boldsymbol{\mathcal{F}}^O_{i_1\ldots i_N} + \frac{n}{n+1}\bar{\boldsymbol{\mathcal{F}}}_{\Omega_{i_1\ldots i_N}}. \tag{A.22}$$

According to (A.18) and (A.22), a conclusion is:

$$(1 - \frac{2n\rho_2}{\mathbb{E}_q[\gamma_{i_1\ldots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2}) = \frac{1}{n+1},$$

and it is same as:

$$\frac{\mathbb{E}_q[\gamma_{i_1\ldots i_N}] + \mathbb{E}_q[\tau]}{\mathbb{E}_q[\gamma_{i_1\ldots i_N}] + \mathbb{E}_q[\tau] + 2n\rho_2} = \frac{1}{n+1}. \tag{A.23}$$

From (A.20), (A.23) can be re-written as:

$$\begin{aligned}
\frac{(\sigma^O_{i_1\ldots i_N})^{-2}}{(\sigma^O_{i_1\ldots i_N})^{-2} + 2n\rho_2} &= \frac{1}{n+1} \\
(n+1)(\sigma^O_{i_1\ldots i_N})^{-2} &= (\sigma^O_{i_1\ldots i_N})^{-2} + 2n\rho_2 \\
n(\sigma^O_{i_1\ldots i_N})^{-2} &= 2n\rho_2 \\
\rho_2 &= \frac{1}{2(\sigma^O_{i_1\ldots i_N})^2}.
\end{aligned}$$

The above is same as (A.21), so the proof is done.

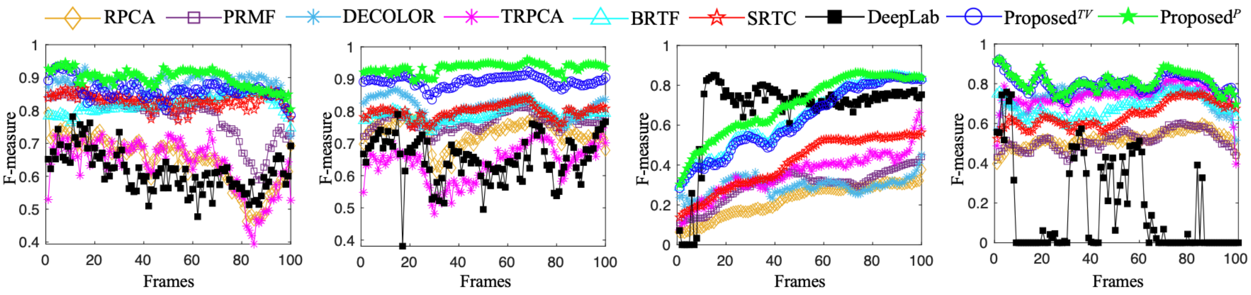
□

Appendix E: Justification for Fig. 3

Fig. 3 in the paper "**Robust Unsupervised Video Foreground Segmentation via Context-Aware Bayesian Tensor Factorization**" illustrates the comparison of the performance consistency in terms of frame-wise F-measure, across 100 frames in each video. The performance of proposed CABTF^{TV}, CABTF^{P}, and all the state-of-the-art unsupervised methods are demonstrated in Fig. 3, where the proposed CABTF^{P} consistently outperforms all other methods with the highest F-measure for almost all frames in each video. The frame-wise F-measure of the supervised DeepLab is not shown in Fig. 3, because it contains zero F-measure that may change the vertical axis of Fig. 3, making the plots of the other methods crowded together and fail to show a clear comparison, as shown in Fig.

A.1(d). The performance comparison including all the unsupervised and supervised methods is illustrated in Fig. A.1. In the video *Pedestrians*, there is a lady walking on the road. The challenges could be overcoming the random camera noise. In the video *Highway*, there are vehicles moving along the highway. The challenges could be focusing on the vehicles without being distracted by the dynamic noise caused by swaying trees on the left side of the highway. In the video *Canoe*, a canoe is moving on a lake and the canoe is filled with people. The challenges could be overcoming the dynamic noise from rippling water. And in the video *Fountain02*, a car with a trail is driving through while being partially blocked the spraying fountain. The challenge could be overcoming the dynamic noise from the fountain and segment the moving vehicle.

Distinct F-measure variation of DeepLab indicates its performance inconsistency, as presented in Fig. A.1. DeepLab is pre-trained on the PASCAL-VOC dataset [3], which consists of more than 10000 images across 21 object classes, including vehicles, pedestrians, and boats, but not including swaying leaves, rippling water, and spraying fountain. Therefore, DeepLab is robust to such dynamic backgrounds, as it cannot recognize them. For unsupervised learning methods, the dynamic backgrounds such as rippling water in *Canoe* severely deteriorate their performance, as shown in the first sixty frames of Fig. A.1(c). However, the DeepLab fails to detect the people sitting in the canoe, and cannot identify the canoe if only part of it is presented in the frames, such as the first ten frames in Fig. A.1(c). Moreover, DeepLab produces considerable zero F-measures for *Fountain02* dataset, as shown in Fig. A.1(d). Because of the small size of the vehicle and potential blurring caused by the fountain, DeepLab fails to identify the vehicle. The limited segmentation accuracy of DeepLab can be potentially caused by the disparity between the training and testing datasets, which is avoidable in complex real-world scenarios. This suggests the limitation of supervised deep learning methods in terms of generalization. Overall, The proposed CABTF^P attains better performance, consistently, than all other methods consistently for almost all frames in each video.



(a) F-measure for *Pedestrians* (b) F-measure for *Highway* (c) F-measure for *Canoe* (d) F-measure for *Fountain02*

Figure A.1: Segmentation performance comparison by frames for different methods in four videos: (a) *Pedestrians*; (b) *Highway*; (c) *Canoe*; (d) *Fountain02*.

References

[1] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[2] Qibin Zhao, Guoxu Zhou, Liqing Zhang, Andrzej Cichocki, and Shun-Ichi Amari. "Bayesian robust tensor factorization for incomplete multiway data". In: *IEEE transactions on neural networks and learning systems* 27.4 (2015), pp. 736–748.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. "The Pascal Visual Object Classes (VOC) Challenge". In: *International Journal of Computer Vision* 88.2 (June 2010), pp. 303–338.